Distributionally Robust Causal Abstractions

Yorgos Felekis[1], **Theodoros Damoulas**[1,2], **Paris Giampouras**[1]

[1]Department of Computer Science, University of Warwick, Coventry, UK
[2]Department of Statistics, University of Warwick, Coventry, UK

{yorgos.felekis, t.damoulas, paris.giampouras}@warwick.ac.uk

Abstract

Causal Abstraction (CA) theory provides a principled framework for relating causal models that describe the same system at different levels of granularity while ensuring interventional consistency between them. Recently, several approaches for learning CAs have been proposed, but all assume fixed and well-specified exogenous distributions, making them vulnerable to environmental shifts and misspecification. In this work, we address these limitations by introducing the first class of distributionally robust CAs and their associated learning algorithms. The latter cast robust causal abstraction learning as a constrained min-max optimization problem with Wasserstein ambiguity sets. We provide theoretical results, for both empirical and Gaussian environments, leading to principled selection of the level of robustness via the radius of these sets. Furthermore, we present empirical evidence across different problems and CA learning methods, demonstrating our framework's robustness not only to environmental shifts but also to structural model and intervention mapping misspecification.

1 Introduction

Causal reasoning provides the foundation for understanding and influencing complex systems: it enables us to move beyond correlation to estimate the effects of interventions, simulate counterfactual scenarios, and uncover the underlying mechanisms that drive a system's behavior. These capabilities are crucial for supporting out-of-distribution generalization and for developing interpretable, fair, and socially aligned machine learning systems. The emerging field of Causal Representation Learning (CRL) [Schölkopf et al., 2021, Locatello et al., 2019] aims to recover structured, latent representations that not only compress data but also retain the causal features necessary for robust reasoning and decision-making. This challenge becomes even more apparent when dealing with systems that naturally operate at multiple levels of abstraction, such as cells versus organs in biology, neurons versus brain regions in neuroscience, or individual behaviors versus population dynamics in social sciences, where causal relationships may manifest differently across scales. While fine-grained models are often assumed to better reflect the underlying dynamics of a system because they are more expressive, they tend to be opaque and computationally expensive. Conversely, more abstract models offer interpretability and efficiency, but are faithful only if they preserve the causal semantics of the systems they simplify. This trade-off motivates the need for consistent causal reasoning *across* abstraction levels in a principled way, allowing for reliable reasoning and decision-making at both detailed and aggregate views of a system.

The theory of Causal Abstractions (CAs) addresses this challenge by formalizing causally consistent models at different granularities and characterizing the abstraction maps that relate them. Two main theoretical frameworks have shaped this space: τ-ω abstractions [Rubenstein et al., 2017, Beckers and Halpern, 2019], which define a single mapping between model domains, and (R, a, α) abstractions [Rischel, 2020], which separate mappings between variables and of their domains. Their relation has been recently analyzed in [Schooltink and Zennaro, 2025]. Recently, *Causal Abstraction Learning (CAL)* [Zennaro et al., 2022], the problem of learning abstraction maps directly from data, has attracted growing attention as it enables cross-scale representation learning, evidence synthesis from heterogeneous sources, and more efficient data collection and modeling.

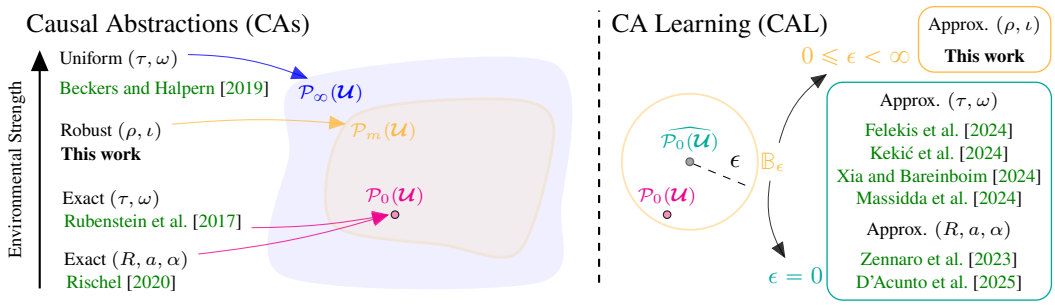


Figure 1: Left: Exact CAs assume a single environment $\mathcal{P}_0(\mathcal{U})$, robust CAs operate over a constrained subset of relevant environments $\mathcal{P}_m(\mathcal{U})$, with $0 \leqslant m < \infty$, and uniform CAs across all possible environments $\mathcal{P}_\infty(\mathcal{U})$. Right: CAL methods in relation to this structure. Our framework models environmental uncertainty using a Wasserstein ball $\mathbb{B}_\epsilon$, centered at the empirical joint environment $\widehat{\mathcal{P}_0(\mathcal{U})}$. Prior methods correspond to $\epsilon = 0$, implicitly assuming a fixed environment.

The first learning method was introduced by Zennaro et al. [2023] on the (R, a, α)-abstraction, followed by several approaches, see Fig. 1, for learning approximate abstractions including clustering-based [D'Acunto et al., 2025, Kekić et al., 2024, Xia and Bareinboim, 2024] and optimal transport ones [Felekis et al., 2024], along with diverse CA and CAL applications, from climate modeling [Chalupka et al., 2017] to neural network interpretability [Geiger et al., 2021], agent-based modeling [Dyer et al., 2024], multi-armed bandits [Zennaro et al., 2024], and causal discovery [Massidda et al., 2024]. A key challenge in CAL is understanding how the environmental conditions, i.e. the exogenous distributions under which an abstraction is learned, influence its generalization and causal consistency properties. While exact abstractions Rubenstein et al. [2017], Rischel [2020] assume a fixed environment, without explicitly addressing how the abstraction depends on it, Beckers and Halpern [2019] propose a notion of abstraction uniformity that requires causal consistency to hold across *all*, infinitely many, environments, see Fig. 1. The latter corresponds to an environment-independent mapping that captures a more fundamental relationship between models, that of their respective deterministic parts, i.e. their causal bases (see Def. 2.1). Although theoretically stronger, this notion of abstraction is computationally infeasible to learn, as it requires data from infinitely many environments including ones that may be unattainable due to physical or practical constraints.

In this work, we address this challenge by introducing a framework of *robust causal abstractions*, which requires causal consistency to hold not uniformly, but rather across a constrained set of relevant environments. This yields a notion that is theoretically stronger than exact abstractions, capturing relationships at a more fundamental level, while remaining more flexible and computationally tractable than uniform abstractions. Building on this, we develop CAL algorithms that model uncertainty over the environments and learn robust CAs that remain consistent under distributional shifts and environmental misspecification. Moreover, as we demonstrate, this robustness extends to other forms of misspecification, including structural assumptions of the causal models and mapping of their respective interventions. In the finite sample case, our approach links the level of modeled uncertainty to the strength of the learned abstraction: the broader the ambiguity set, the more powerful the abstraction it approximates. Overall, we make the following contributions:

- We introduce the first class of *robust causal abstractions*, called (ρ, ι)-abstractions, that require consistency across a constrained set of environments. This yields stronger abstractions than exact ones (fixed environment) yet more realistic and tractable than uniform ones (all environments).
- We introduce *Distributionally Robust Causal Abstractions* (DiRoCA), the first framework for learning CAs robust to distributional shifts and misspecifications in the underlying SCM environments. Our approach formulates learning as a distributionally robust optimization (DRO) problem over a Wasserstein ambiguity set, explicitly modeling environment uncertainty.
- We provide theoretical results for joint product measures under both Gaussian and empirical settings, to guide the selection of the robustness parameter in our DRO setting.
- We demonstrate the effectiveness of DiRoCA across different problems and scenarios, consistently outperforming existing methods and baselines under environmental shifts, structural and intervention mapping misspecifications.

2 Background

Causal Models, Environments and Abstractions. We work with Pearl [2009]'s Structural Causal Models, where each endogenous variable is a function of its direct causes and the exogenous noise.

Definition 2.1 (Structural Causal Model)**.** A d-dimensional *Structural Causal Models (SCM)* is a pair $\mathcal{M}^d \coloneqq (\mathcal{S}^d, \rho^d)$, where $\mathcal{S}^d = \langle \mathcal{X}, \mathcal{U}, \mathcal{F} \rangle$ defines the deterministic *causal basis*, consisting of a set of d *endogenous variables* $\mathcal{X}$, a set of d *exogenous variables* $\mathcal{U}$ and a set of d *structural functions* $\mathcal{F}$, each defining the value of an endogenous variable as $X_i = f_i(\mathrm{PA}(X_i), U_i) \quad \forall\, i \in [d]$, where $\mathrm{PA}(X_i) \subseteq \mathcal{X} \backslash X_i$ denotes the set of parent variables (direct causes) of X_i. The measure ρ^d, is a joint probability distribution over the exogenous variables $\mathcal{U}$, called the **environment** of the SCM.

Causal Assumptions. We focus on *Markovian* SCMs, where the exogenous variables are jointly independent; i.e. $\rho^d = \prod_{i=1}^{d} \mathbb{P}(U_i)$, a property that entails several standard assumptions. Markovianity implies *acyclicity*, so $\mathcal{M}^d$ entails a directed acyclic graph (DAG) $\mathcal{G}_{\mathcal{M}^d}$ where the nodes correspond to the endogenous variables $\mathcal{X}$ and the edges are defined by the signature of the functions f_i in $\mathcal{F}$. We assume this graph is given. In addition, due to acyclicity, the structural functions $\mathcal{F}$ can be recursively composed into a single deterministic map $\mathbf{g} : \mathrm{dom}[\mathcal{U}] \to \mathrm{dom}[\mathcal{X}]$, referred to as the *mixing function* of the model. This defines the SCM's *reduced form*, in which the endogenous variables are expressed purely as a function of the exogenous variables; i.e. $\mathcal{X} = \mathbf{g}(\mathcal{U})$. The induced distribution over $\mathcal{X}$ can then be expressed as: $\mathbb{P}_{\mathcal{M}^d}(\mathcal{X}) = \mathbf{g}_{\#}(\rho^d)$, where $\mathbf{g}_{\#}$ denotes the pushforward measure, making explicit the generative process by which exogenous uncertainty propagates through the causal model. It also implies *causal sufficiency*, meaning that there are no unobserved confounders. Furthermore, we assume *faithfulness*, meaning that independencies in the data are captured in the graphical model. Finally, we focus on *Linear Additive Noise (LAN)* models [Peters et al., 2016], a class of SCMs where the endogenous variables satisfy $\mathcal{X} = \mathbf{B}^\top \mathcal{X} + \mathcal{U}$, with $\mathbf{B} \in \mathbb{R}^{d \times d}$ the weighted adjacency matrix of the causal DAG $\mathcal{G}_{\mathcal{M}^d}$ and $\mathcal{U}$ a vector of independent exogenous noise variables. Due to acyclicity, $\mathbf{B}$ can be permuted to an upper triangular form, yielding the reduced form $\mathcal{X} = \mathbf{M}\,\mathcal{U},$ with $\mathbf{M} = (\mathbb{I}^d - \mathbf{B})^{-1}$. The linear transformation, $\mathbf{M}$ is the *mixing matrix*, which linearly maps the independent exogenous noise to the observed endogenous data.

Interventions. A key feature of SCMs is their ability to facilitate reasoning about *interventions* in a causal system. For instance, an intervention $\iota = \mathrm{do}(\mathcal{A} = \mathbf{a})$ sets each variable A_i in $\mathcal{A} \subseteq \mathcal{X}$ to a specific value a_i, while allowing the rest of the system to evolve as usual. Graphically, such an operation *mutilates* the DAG $\mathcal{G}_{\mathcal{M}^d}$ by removing the incoming edges to variables in $\mathcal{A}$, yielding a new post-interventional SCM $\mathcal{M}^d_\iota$ and thus altering the distribution of the remaining variables, giving rise to the *interventional distribution* $\mathbb{P}_{\mathcal{M}^d_\iota}(\mathcal{X}) = \mathbb{P}(\mathcal{X} \mid \mathrm{do}(\mathcal{A} = \mathbf{a}))$. These are called *exact* interventions, as they deterministically set a variable to a fixed value, fully overriding its causal dependencies, leaving no residual influence.

Causal Abstractions. Causal Abstraction (CA) theory formalizes the relation between SCMs defined at different levels of granularity by requiring *interventional consistency* between them. This enables flexible shifts in the level of representation used for causal reasoning, depending on the specific inquiry or the nature of the available data.

Definition 2.2 ([Rubenstein et al., 2017])**.** Let SCMs $\mathcal{M}^\ell$ and $\mathcal{M}^h$ with fixed respective environments ρ^ℓ, ρ^h and intervention sets $\mathcal{I}^\ell$, $\mathcal{I}^h$ and a surjective and order-preserving map $\omega : \mathcal{I}^\ell \to \mathcal{I}^h$. An abstraction $\tau : \mathrm{dom}[\mathcal{X}^\ell] \to \mathrm{dom}[\mathcal{X}^h]$, with $\ell \geqslant h$, is called an *exact transformation* if:

$$\tau_{\#}(\mathbb{P}_{\mathcal{M}^\ell_\iota}(\mathcal{X}^\ell)) = \mathbb{P}_{\mathcal{M}^h_{\omega(\iota)}}(\mathcal{X}^h), \quad \forall\, \iota \in \mathcal{I}^\ell \tag{1}$$

This is illustrated in Fig. 2 by comparing the green (intervene–then–abstract) and purple (abstract–then–intervene) paths; their mismatch defines what is known as the *abstraction error*. We further formalize these concepts in the next section.

3 Distributionally Robust Causal Abstractions

In this section, we introduce a new class of CAs that extends exact transformations by ensuring interventional consistency of the two models across a set of environments, rather than just one. This formulation yields a dual benefit: it defines a stronger notion of abstraction, and it enables robustness to misspecification and distributional shifts when learning the abstraction from data.

The (ρ, ι)-Causal Abstraction Framework. We explicitly model the uncertainty inherent in each SCM through its corresponding environment: ρ^ℓ for the low-level model $\mathcal{M}^\ell$ and ρ^h for the high-level model $\mathcal{M}^h$. These give rise to a *joint environment* $\boldsymbol{\rho} = \rho^\ell \otimes \rho^h$, a product measure that captures the uncertainty across abstraction

levels. Formally, $\boldsymbol{\rho} \in \mathcal{P}(\boldsymbol{\mathcal{U}})$, where $\boldsymbol{\mathcal{U}} = \mathrm{dom}[\mathcal{U}^\ell] \times \mathrm{dom}[\mathcal{U}^h]$ denotes the product measure space of the exogenous domains. In practice, the realization of certain environments is infeasible or unrealistic. Hence, we introduce the notion of *relevant environments*, denoted $\mathcal{A}^\ell \subseteq \mathcal{P}(\mathcal{U}^\ell)$ and $\mathcal{A}^h \subseteq \mathcal{P}(\mathcal{U}^h)$ for the low- and high-level SCMs, respectively. Accordingly, we define the *relevant joint environment space* as $\mathcal{A} = \mathcal{A}^\ell \otimes \mathcal{A}^h$, which represents all joint distributions formed by pairing any $\rho^\ell \in \mathcal{A}^\ell$ with any $\rho^h \in \mathcal{A}^h$. This mirrors the concept of *relevant interventions* $\mathcal{I}$ that is a partially ordered set that restricts attention to interventions that are semantically meaningful or practically implementable. Furthermore, to align interventions across abstraction levels, we assume the existence of a surjective and order-preserving map $\omega : \mathcal{I}^\ell \to \mathcal{I}^h$. To ensure consistent alignment of joint interventions across levels, we define $\mathcal{I} = \{(\iota, \omega(\iota)) \mid \iota \in \mathcal{I}^\ell\}$, pairing each low-level intervention with its high-level counterpart. Finally, we define the *abstraction context* as the pair $(\mathcal{A}, \mathcal{I})$, which specifies the set of interventions and environments over which an abstraction is evaluated. Building on this setup, we introduce a refined notion of abstraction that operates on the reduced form of the SCM, making explicit the dependence of the abstraction on both the environment spaces and the intervention sets associated with the low- and high-level models.

Definition 3.1 ((ρ, ι)-Abstraction)**.** Let $(\mathcal{A}, \mathcal{I})$ be an abstraction context, where $\mathcal{M}^\ell = (\mathcal{S}^\ell, \rho^\ell)$, $\mathcal{M}^h = (\mathcal{S}^h, \rho^h)$ the low-level and high-level SCMs with $\ell \geqslant h$, and let $\mathbf{g}^\ell$ and $\mathbf{g}^h$ denote the respective mixing functions of their reduced forms. Then a map $\tau : \mathrm{dom}[\mathcal{X}^\ell] \to \mathrm{dom}[\mathcal{X}^h]$ is called a (ρ, ι)-*abstraction* if, for all $\boldsymbol{\rho} = \rho^\ell \otimes \rho^h \in \mathcal{A}$ and for all $\boldsymbol{\iota} = (\iota, \omega(\iota)) \in \mathcal{I}$:

$$\tau_{\#}\left(\mathbf{g}^\ell_{\iota\#}(\rho^\ell)\right) = \mathbf{g}^h_{\omega(\iota)\#}(\rho^h) \tag{2}$$

A (ρ, ι)-abstraction ensures commutativity between interventions and abstractions under a given joint environment $\boldsymbol{\rho}$: **(a)** intervening via ι and then abstracting via τ yields the same result as **(b)** abstracting first and then intervening via $\omega(\iota)$ for all $\iota \in \mathcal{I}^\ell$. In general, we say that $\mathcal{M}^h$ and $\mathcal{M}^\ell$ are (ρ, ι)-*interventionally consistent* if there exists an abstraction context $(\mathcal{A}, \mathcal{I})$ such that Eq. 2 holds. Furthermore, in this work, we focus on the case of *linear abstractions* [Massidda et al., 2024], an important and well-behaved class of abstractions according to which the abstraction map can be represented through a matrix $\mathrm{T} \in \mathbb{R}^{h \times \ell}$ and thus the map τ can be represented as a matrix-vector multiplication $\tau(x) = \mathrm{T}\, x$. Let $\mathbf{L}_\iota$ and $\mathbf{H}_{\omega(\iota)}$ be the linear transformations of the reduced forms $\mathbf{g}^\ell$ of $\mathcal{M}^\ell$ and $\mathbf{g}^h$ of $\mathcal{M}^h$ for every intervention, then Eq. 2 becomes: $\mathrm{T}\, \mathbf{L}_\iota \mathcal{U}^\ell = \mathbf{H}_{\omega(\iota)} \mathcal{U}^h$.

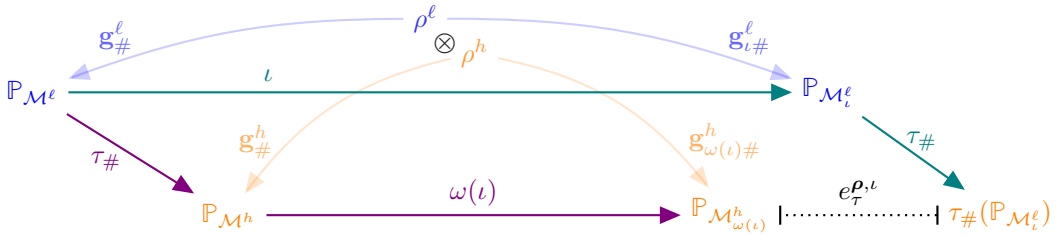


Figure 2: *Computation of the environment–intervention error.* The joint environment $\boldsymbol{\rho} = \rho^\ell \otimes \rho^h$ captures the combined uncertainty from the low- and high-level SCMs. By pushing forward these components through the reduced forms $\mathbf{g}^\ell$ and $\mathbf{g}^h$ of the respective SCMs, we evaluate two interventional pathways: (a) apply an intervention ι to $\mathcal{M}^\ell$, then map the resulting distribution to the high-level space via $\tau_{\#}$: $\tau_{\#}(\mathbb{P}_{\mathcal{M}^\ell_\iota})$; and (b) first map $\mathcal{M}^\ell$ to $\mathcal{M}^h$ via $\tau_{\#}$, then apply the corresponding intervention $\omega(\iota)$: $\mathbb{P}_{\mathcal{M}^h_{\omega(\iota)}}$. The divergence $\mathcal{D}_{\mathcal{X}^h}$ between the resulting interventional distributions computes $e^{\boldsymbol{\rho},\iota}_\tau$. Aggregating $e^{\boldsymbol{\rho},\iota}_\tau$ over an $(\mathcal{A}, \mathcal{I})$ abstraction context, recovers the $(\mathcal{A}, \mathcal{I})$–abstraction error (Eq. 3). If this is zero, the diagram commutes and τ defines a τ–0–approximate abstraction.

Our (ρ, ι)-framework generalizes and bridges prior notions of causal abstraction. Unlike exact transformations [Rubenstein et al., 2017], which assume consistency under some fixed (but unspecified) environment, and uniform transformations [Beckers and Halpern, 2019], which require consistency across all environments, (ρ, ι) allows for a range of finite, plausible environments. This enables principled abstraction under realistic data constraints. Further discussion on this is in Appendix A.1.

The $(\mathcal{A}, \mathcal{I})$–Abstraction Error. Def. 3.1 suggests a perfect form of abstraction where interventional consistency holds exactly, implying that under a given abstraction context $(\mathcal{A}, \mathcal{I})$ using the high-level model entails no loss in predictive accuracy. However, in reality, such types of abstractions are rare. This stems in part from the very nature of abstraction, which reduces the size of the representation by disregarding minor differences, often leading to some degree of information loss. More importantly, in CAL, where the goal is to learn abstractions from data, all

inferences are inherently approximate. Thus we need a notion of *abstraction error* [Beckers et al., 2020] tailored to our framework. The core idea is to measure the discrepancy between the two sides of Eq. 2 while accounting for the given context $(\mathcal{A}, \mathcal{I})$. To this end, we assume access to a metric over high-level interventional distributions, which induces a discrepancy between SCMs via τ enabling approximate abstractions. Specifically, we introduce a novel notion of abstraction error by aggregating over environments and interventions.

Definition 3.2. Let $\tau : \mathrm{dom}[\mathcal{X}^\ell] \to \mathrm{dom}[\mathcal{X}^h]$ be a map between the SCMs $\mathcal{M}^\ell$ and $\mathcal{M}^h$, defined wrt an abstraction context $(\mathcal{A}, \mathcal{I})$. Let also $D_{\mathcal{X}^h}$ be a discrepancy measure over high-level distributions, and q be a distribution over interventions in $\mathcal{I}^\ell$. We define the $(\mathcal{A}, \mathcal{I})$-*Abstraction Error* as:

$$e_\tau(\mathcal{M}^\ell, \mathcal{M}^h) = \mathop{\mathfrak{g}}_{\boldsymbol{\rho} \in \mathcal{A}} \mathop{\mathfrak{h}}_{\iota \sim q} \left[e_\tau^{\boldsymbol{\rho}, \iota}(\mathcal{M}^\ell, \mathcal{M}^h) \right], \tag{3}$$

where we define $e_\tau^{\boldsymbol{\rho}, \iota}(\mathcal{M}^\ell, \mathcal{M}^h) = \mathcal{D}_{\mathcal{X}^h}\left(\tau_\#(\mathbf{g}^\ell_{\iota\#}(\rho^\ell)),\ \mathbf{g}^h_{\omega(\iota)\#}(\rho^h) \right)$ as the *environment–intervention* approximation error induced by τ for a fixed $\boldsymbol{\rho} \in \mathcal{A}$ and intervention $\iota \in \mathcal{I}^\ell$.

The operators $\mathfrak{g}$ and $\mathfrak{h}$ aggregate the error over environments and interventions, respectively. Typically, prior work has focused on a sub-case of Eq.3 where aggregation is performed only over interventions ($\mathfrak{h}$) either as an expectation [Felekis et al., 2024, Dyer et al., 2024, Kekić et al., 2024] or as a supremum [Zennaro et al., 2023], assuming a fixed environment. [Beckers et al., 2020] proposes a similar notion of abstraction error, though their framework is defined under uniform abstractions and does not consider restrictions to subsets of the environment space, as we do. We say that $\mathcal{M}^h$ is a *approximate* (ρ, ι)*-abstraction* of $\mathcal{M}^\ell$ if $e_\tau(\mathcal{M}^\ell, \mathcal{M}^h) > 0$. A visual representation of this approximation is illustrated in Fig. 2. The connection between Def. 3.2 and (ρ, ι)-abstractions now becomes apparent. In particular, when the total abstraction error vanishes for a given τ, the pushforward of the low-level interventional distributions $\mathbf{g}^\ell_{\iota\#}(\rho^\ell)$ through $\tau_\#$ matches the corresponding high-level ones $\mathbf{g}^h_{\omega(\iota)\#}$, q-almost surely. Statement and proof of this in Proposition 1, Appendix A.4.

4 Learning Distributionally Robust Causal Abstractions

We learn *robust causal abstractions* when data is available from *a single environment per SCM*. Consider a pair $(\mathcal{M}^\ell, \mathcal{M}^h)$ of SCMs and define their joint environment $\boldsymbol{\rho}$ as before, representing the unknown data-generating process. We learn an abstraction map τ that reliably transforms low-level distributions into high-level ones, even in the presence of distributional shifts or test-time noise. To address this, we build on Wasserstein *Distributionally Robust Optimization (DRO)* [Kuhn et al., 2019], which provides robustness against distributional misspecification. The standard objective is:

$$\inf_{x \in \mathcal{X}} \sup_{\mathbb{Q} \in \mathbb{B}_{\varepsilon,p}(\widehat{\mathbb{P}}_N)} \mathbb{E}_{\xi \sim \mathbb{Q}}[f(x, \xi)], \tag{4}$$

The goal is to minimize the worst-case expected loss $f(x, \xi) : \mathbb{R}^n \times \Xi \to \mathbb{R}$, where $x \in \mathbb{R}^n$ denotes a decision variable, $\xi \in \Xi = \mathbb{R}^m$ a random vector representing uncertain data, over a set of plausible distributions, called the *ambiguity set*. In our setting, the loss is the *environment-intervention error* $e_\tau^{\boldsymbol{\rho}, \iota}(\mathcal{M}^\ell, \mathcal{M}^h)$, the decision variable is τ, and we minimize the worst-case abstraction error over a *2-Wasserstein product ambiguity set* centered at the empirical joint environment $\widehat{\boldsymbol{\rho}}$ with radius ϵ. More details on DRO are provided in Appendix A.3.

Environmental Robustness via (ρ, ι)-Abstractions. We introduce distributional robustness by setting the aggregation function $\mathfrak{g}$ in Eq. 3 over the environments to a supremum, and for computational practicality, the aggregation function $\mathfrak{h}$ over the interventions to an expectation. Following the DRO paradigm, and since we assume access to a single environment from each SCM, we substitute $\mathcal{P}(\mathcal{A})$ with a *joint ambiguity set* over the low- and high-level environments to incorporate distributional uncertainty in a principled way using a 2-Wasserstein product ball centered at the empirical joint environment $\widehat{\boldsymbol{\rho}}$. We denote this ambiguity set by $\mathbb{B}_{\epsilon,2}(\widehat{\boldsymbol{\rho}})$, which contains all the environments that lie within a radius ϵ, determined by the individual radii ε_ℓ and ε_h, from the nominal empirical components. Formally, it is given by:

$$\mathbb{B}_{\epsilon,2}(\widehat{\boldsymbol{\rho}}) := \mathbb{B}_{\varepsilon_\ell,2}(\widehat{\rho^\ell}) \times \mathbb{B}_{\varepsilon_h,2}(\widehat{\rho^h}), \tag{5}$$

where $\mathbb{B}_{\varepsilon_d,2}(\widehat{\rho^d}) = \left\{ \rho \in \mathcal{P}(\mathcal{U}^d) : W_2(\rho, \widehat{\rho^d}) \leqslant \varepsilon_d \right\}$ for $d \in \{\ell, h\}$, is a 2-Wasserstein ball centered at $\widehat{\rho^d}$. The radius ε_d controls the robustness level: larger values provide protection against broader shifts but may yield more conservative solutions. Each marginal Wasserstein ball bounds the allowable deviation from the empirical environment for each SCM. As we have sample access to the endogenous variables, we estimate the exogenous

environments via abduction by inverting the reduced form of the SCM, denoted as $(\mathbf{g}^d_{\#})^{-1}$. In the case of LAN models, if the structural functions are known, this inversion is exact; else, they must be inferred, and the abduction step recovers an approximate exogenous distribution. The role of the joint ambiguity set $\mathbb{B}_{\epsilon,2}(\widehat{\boldsymbol{\rho}})$ is to identify the worst-case joint environment $\boldsymbol{\rho}^\star = \rho^{\ell\star} \otimes \rho^{h\star}$ within the product ball that maximizes the abstraction error. This captures the most adversarial shift consistent with the estimation error. The full construction process is illustrated in Fig. 5 of Appendix A.5 for the case of independently sampled environments.

The DIRoCA Objective. Our objective is to learn a *linear abstraction map* τ, that remains reliable across all shifts inside the 2-Wasserstein product ambiguity set $\mathbb{B}_{\epsilon,p}(\widehat{\boldsymbol{\rho}})$. Assuming access to an abstraction context $(\mathcal{A}, \mathcal{I})$ and a divergence measure $\mathcal{D}_{\mathcal{X}^h}$ over high-level interventional distributions, we define the *distributionally robust causal abstraction learning objective* as follows:

$$\min_{\mathrm{T}\in\mathbb{R}^{h\times\ell}} \sup_{\boldsymbol{\rho}\in\mathbb{B}_{\epsilon,p}(\widehat{\boldsymbol{\rho}})} \mathop{\mathbb{E}}_{\iota\sim q}\left[e^{\boldsymbol{\rho},\iota}_{\tau}(\mathcal{M}^\ell, \mathcal{M}^h)\right] \tag{6}$$

where $\mathrm{T} \in \mathbb{R}^{h\times\ell}$ denotes the class of linear abstraction maps τ, and $e^{\boldsymbol{\rho},\iota}_{\tau}(\cdot,\cdot)$ is the environment–intervention error. This objective seeks the linear abstraction map that minimizes the worst-case expected abstraction error over the joint ambiguity set, ensuring robustness to distributional shifts.

Remark 1. *The robustness radius ϵ provides a natural interpolation between prior causal abstraction frameworks. As $\epsilon \to 0$, the DIRoCA objective recovers a (τ,ω)-transformation, optimized solely for the empirical environment $\widehat{\boldsymbol{\rho}}$. Conversely, as $\epsilon \to \infty$, DIRoCA seeks robustness over all possible environments, resulting in a stronger abstraction and approximates a uniform transformation. Thus, both (τ,ω) and uniform transformations arise as special cases of our framework, which flexibly interpolates between these extremes depending on the desired level of robustness.*

Concentration guarantees for the joint environment. Our theoretical results below show that in both a Gaussian and empirical setting, the true joint environment $\boldsymbol{\rho}$ lies within a 2-Wasserstein product ball centered at $\widehat{\boldsymbol{\rho}}$ with high probability for an appropriate radius ϵ. *This result offers a principled way to set the robustness parameter ϵ to ensure the ambiguity set reliably covers the true environment*, thus justifying our distributionally robust CA objective for independent environments.

Assumption 1. For $d \in \{\ell, h\}$, ρ^d is a light-tailed environment; i.e. there exist constants $\alpha > 0$ and $A > 0$ such that $\mathbb{E}_{\rho^d}\left[\exp\left(\|\xi\|_2^\alpha\right)\right] \leqslant A$, where $\xi \sim \rho^d$.

Theorem 1 (Gaussian $\boldsymbol{\rho}$-Concentration)**.** *Let $\rho^\ell \sim \mathcal{N}(\mu_\ell, \Sigma_\ell)$ and $\rho^h \sim \mathcal{N}(\mu_h, \Sigma_h)$, under Assumption 1, let $\widehat{\rho^\ell}$ and $\widehat{\rho^h}$ from N_ℓ and N_h i.i.d. samples. Also, let $\boldsymbol{\rho} := \rho^\ell \otimes \rho^h$ and $\widehat{\boldsymbol{\rho}} := \widehat{\rho^\ell} \otimes \widehat{\rho^h}$. Then, for $d \in \{\ell, h\}$, there exist constants $c_d > 0$, depending only on the individual d-environment and confidence levels η_d, such that $\forall \delta \in (0,1]$, with $\delta = 1 - (1-\eta_\ell)(1-\eta_h)$:*

$$\mathbb{P}\left[\mathcal{W}_2(\boldsymbol{\rho}, \widehat{\boldsymbol{\rho}}) \leqslant \epsilon\right] \geqslant 1-\delta, \;\; \forall\, \epsilon \geqslant \sqrt{\left(\frac{\log(c_\ell/\eta_\ell)}{\sqrt{N_\ell}}\right)^2 + \left(\frac{\log(c_h/\eta_h)}{\sqrt{N_h}}\right)^2}$$

Theorem 2 (Empirical $\boldsymbol{\rho}$-Concentration)**.** *Let $\widehat{\rho^\ell}$ and $\widehat{\rho^h}$ empirical distributions, under Assumption 1, from N_ℓ and N_h i.i.d. samples, with $\boldsymbol{\rho} := \rho^\ell \otimes \rho^h$, $\widehat{\boldsymbol{\rho}} := \widehat{\rho^\ell} \otimes \widehat{\rho^h}$. Then, for $d \in \{\ell, h\}$, there exist constants $c_{d,1}, c_{d,2} > 0$, depending only on the individual d-environment and confidence levels η_d, such that $\forall \delta \in (0,1]$, with $\delta = 1-(1-\eta_\ell)(1-\eta_h)$, for $N_d(c,\eta) = \log(c_{d,1}/\eta)/c_{d,2}$, if we set:*

$$\tilde{\varepsilon_d} = \left(\frac{\log(c_{d,1}/\eta)}{c_{d,2}N_d}\right)^{\min\{1/d,1/2\}} \quad \text{if } N_d \geqslant N_d(c,\eta), \quad \text{or} \quad \left(\frac{\log(c_{d,1}/\eta)}{c_{d,2}N_d}\right)^{1/\alpha_d} \quad \text{otherwise,}$$

$$\implies \quad \mathbb{P}\left[\mathcal{W}_2(\boldsymbol{\rho}, \widehat{\boldsymbol{\rho}}) \leqslant \epsilon\right] \geqslant 1-\delta, \quad \forall\, \epsilon \geqslant \sqrt{\tilde{\varepsilon_\ell}^2 + \tilde{\varepsilon_h}^2}$$

These concentration bounds for CA show how the robustness radius ϵ contracts with sample size to ensure high probability coverage of the true environment. In particular, $\epsilon = \mathcal{O}\left(N^{-1/d}\right)$ for dimension $d > 2$, where $N = \min(N_\ell, N_h)$ and $d = \max(\ell, h)$. Under finite sample, setting $\epsilon > 0$ yields a stronger abstraction, valid over the entire Wasserstein ball. Proofs in Appendix A.4.

Remark 2. *These results assume known coefficients for exact abduction of $\widehat{\boldsymbol{\rho}}$. When these are estimated, the nominal environment may deviate from the true; nevertheless, as the sample size increases, the approximation improves, and the concentration bounds remain asymptotically valid.*

Optimization. We now demonstrate how to solve the robust causal abstraction problem under the model-based (*Gaussian*) and model-free (*empirical*) formulations introduced earlier. In both cases, the problem is cast as a min–max optimization between the abstraction map τ and the worst-case distributions over a Wasserstein ambiguity set, adapted to the available information. In the *Gaussian setting*, leveraging the fact that reduced-form solutions in Linear Additive Noise models (LANs) preserve Gaussianity under linear transformations, the pushforward distributions remain Gaussians under T. The DRO objective then admits a closed-form expression using the 2-Wasserstein distance between Gaussians, known as the Gelbrich distance: $\|\mu_1 - \mu_2\|^2 + \mathrm{Tr}(\Sigma_1) + \mathrm{Tr}(\Sigma_2) - 2\mathrm{Tr}\left((\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\right)$, and the objective becomes:

$$F(\mathrm{T}, \boldsymbol{\rho}) := \mathbb{E}_{\iota\sim q}\left[\mathcal{W}_2^2\left(\mathcal{N}(\mathrm{T}\,\mathbf{L}_\iota\mu^\ell,\, \mathrm{T}\,\mathbf{L}_\iota\Sigma^\ell\mathbf{L}_\iota^\top\,\mathrm{T}^\top),\, \mathcal{N}(\mathbf{H}_{\omega(\iota)}\mu^h,\, \mathbf{H}_{\omega(\iota)}\Sigma^h\mathbf{H}_{\omega(\iota)}^\top)\right)\right], \tag{7}$$

where $\mathcal{N}(\mu^\ell, \Sigma^\ell) \sim \rho^\ell$ and $\mathcal{N}(\mu^h, \Sigma^h) \sim \rho^h$ are the low- and high-level marginals of the joint environment $\boldsymbol{\rho} = \rho^\ell \otimes \rho^h$, with $\boldsymbol{\rho} \in \mathbb{B}_{\epsilon,2}(\widehat{\boldsymbol{\rho}})$. We relax the non-smooth term in the Gelbrich distance using a variational upper bound (see Appendix A.8) and optimize the resulting objective via a proximal gradient ascent–descent scheme. A visual comparison of the initial empirical distributions and the final worst-case distributions found by our algorithm is provided in Appendix A.8 (Figures 18a and 18b). In the *empirical setting*, where only samples are available, we define the robust objective via perturbed empirical distributions. Specifically, we introduce perturbation matrices $\Theta_d \in \mathbb{R}^{N_d\times d}$ over the empirical exogenous samples and define $\rho^d(\Theta_d) = \frac{1}{N_d}\sum_{i=1}^{N_d}\delta_{\hat{u}_i+\theta_i}$. The objective is the squared Frobenius discrepancy between the perturbed distributions of the SCMs:

$$F(\mathrm{T}, \boldsymbol{\rho}) := \mathbb{E}_{\iota\sim q}\left[\left\|\mathrm{T}\;\mathbf{L}_\iota(\mathrm{U}^\ell + \Theta_\ell) - \mathbf{H}_{\omega(\iota)}(\mathrm{U}^h + \Theta_h)\right\|_{\mathrm{F}}^2\right] \tag{8}$$

The perturbations capture adversarial shifts around the nominal distributions and a reparameterization of the Wasserstein ambiguity sets reduces them to constrained Frobenius norm balls [Kuhn et al., 2019]: $\|\Theta_d\|_{\mathrm{F}} \leqslant \rho_d\sqrt{N_d}$. We solve this via alternating projected gradient descent. The general optimization procedure can be seen in Alg. 1 and applies whether structural functions are known or not, with the latter requiring an additional estimation step via linear regression. Further details in Appendix A.8.

Algorithm 1 DIROCA. See Alg. 2 and Alg. 3 for Gaussian and empirical instantiations.

Input: $\mathcal{I}^\ell, \mathcal{I}^\ell, \omega, \mathbb{P}_{\mathcal{M}_\iota^\ell}(\mathcal{X}^\ell), \mathbb{P}_{\mathcal{M}_{\omega(\iota)}^h}(\mathcal{X}^h),\ \forall \iota \in \mathcal{I}^\ell$
$\widehat{\rho^\ell}, \widehat{\rho^h} \leftarrow (\mathbf{g}_\#^\ell(\mathbb{P}_{\mathcal{M}^\ell}))^{-1}, (\mathbf{g}_\#^h(\mathbb{P}_{\mathcal{M}^h}))^{-1}$ ▷ Get environments via abduction
$\mathbb{B}_{\epsilon,p}(\widehat{\boldsymbol{\rho}}) \leftarrow \widehat{\boldsymbol{\rho}} = \rho^\ell \otimes \rho^h$ and ϵ from Theorem 1, 2
Initialize: $\mathrm{T}^{(0)}$, $\rho^{(0)}$, $F(\cdot,\cdot)$ from Eq. 7, 8
repeat until convergence:
  $\rho^{(t+1)} \leftarrow \arg\max_{\rho\in\mathbb{B}_{\epsilon,p}(\widehat{\boldsymbol{\rho}})} F(\mathrm{T}^{(t)}, \rho)$ ▷ Update worst-case environment
  $\mathrm{T}^{(t+1)} \leftarrow \arg\min_\tau F(\mathrm{T}, \rho^{(t+1)})$ ▷ Update abstraction map
return: $\mathrm{T}^\star \in \mathbb{R}^{h\times\ell}, \rho^\star$

5 Experimental Results

We examine DIROCA's performance versus prior art across problems, misspecification levels, and distributional shifts. Further results, evaluation details, and a roadmap of experiments can be found in Appendix A.7. The code for all experiments is publicly accessible[1].

Overview. We denote our method as $\text{DIROCA}_{\varepsilon_\ell,\varepsilon_h}$ to emphasize the robustness parameters that define the radius ϵ of the joint ambiguity set that was used during training. Also, $\text{DIROCA}_{\varepsilon_\ell^\star,\varepsilon_h^\star}$ denotes the method trained with robustness radii set to the theoretical lower bounds derived from Theorems 1, 2. We evaluate its performance across Gaussian and empirical environments based on the empirical approximation of the $(\mathcal{A},\mathcal{I})$-abstraction error from Eq. (3) for $\mathfrak{g}, \mathfrak{h}$ both instantiated as the expected value ($\mathbb{E}$), the distance metric $\mathcal{D}_{\mathcal{X}^h}$ is the 2-Wasserstein distance for the Gaussian setting or the Frobenius distance for the empirical setting, and report $\mu \pm \sigma$ abstraction error ($p \geqslant 0.05$, paired t-test). For all experiments, we evaluate performance using a k-fold cross-validation scheme (with $k = 5$). We assume access to the structural functions and the sets of discrete interventions for both the low- and high-level models alongside their mapping ω.

[1] DIROCA public codebase

Data and Methods. We work with the following datasets: **(a)** *Simple Lung Cancer (SLC)*: A low-level SCM with a causal chain structure ($\ell = 3$) in which the abstracted model ($h = 2$) removes a mediator node and *(b) Linear LiLUCAS*: A linearized version of the LUng CAncer Simple, a higher dimensional benchmark ($\ell = 6, h = 3$) that simulates realistic causal relationships in lung cancer diagnosis. A complete description of them can be found in Appendix A.6. Furthermore, we compare DIROCA against the following methods: **(a)** $\text{BARY}_{(\tau,\omega)}$, a variation of the BARY_{OT} of [Felekis et al., 2024] which computes the Wasserstein barycenters of the interventional distributions at both abstraction levels and learns a mapping between them; **(b)** $\text{GRAD}_{(\tau,\omega)}$ that ignores environmental uncertainty and approximates a (τ, ω) transformation via gradient descent; corresponding to $\text{DIROCA}_{0,0}$ as discussed in Remark 1 and for the empirical setting we also compare against **(c)** ABS-LINGAM [Massidda et al., 2024], that estimates the abstraction map from observational data by solving a least squares problem between low- and high-level samples. We evaluate both variants proposed in the original work: the *perfect* (AbsLin_{p}) and *noisy* (AbsLin_{n}) abstraction learning variants.

Evaluation. We perform a robustness analysis using a unified evaluation framework based on the Huber contamination model. This is applied to the held-out test set for each of the k cross-validation folds, controlled by two key parameters: the contamination fraction $\alpha \in [0, 1]$, which controls the *prevalence* of shifted samples, and the contamination strength $\tilde{\sigma}$, which controls the magnitude of their shift. Each test set consists of a collection of paired, clean endogenous datasets corresponding to each intervention $\{(X_\iota^\ell, X_{\omega(\iota)}^h)\}_{\iota \in \mathcal{I}_\ell} \in \mathbb{R}^{N_{\text{test}}^\ell \times \ell} \times \mathbb{R}^{N_{\text{test}}^h \times h}$. For each intervention ι, we first generate a noisy version of the data, $\tilde{X}_\iota^\ell$ by applying an additive stochastic shift to the clean data. This is achieved by creating a noise matrix $\mathbf{N} \in \mathbb{R}^{N_{\text{test}}^\ell \times \ell}$, where each row $\mathbf{n}_i \sim \mathcal{N}(0, \tilde{\sigma}^2 \cdot \mathbf{I})$ is an independent sample from a zero-mean Gaussian distribution with a scaled covariance and thus, $\tilde{X}_\iota^\ell = X_\iota^\ell + \mathbf{N}$. While we focus on Gaussian shifts in the main text for clarity, we present analogous results for Student-t and Exponential noise distributions in Appendix A.7. The final contaminated test sets, $\bar{X}_\iota^\ell$ for every $\iota \in \mathcal{I}^\ell$, are then formed as a convex combination:

$$\bar{X}_\iota^\ell = (1 - \alpha) X_\iota^\ell + \alpha \tilde{X}_\iota^\ell, \tag{9}$$

An analogous process applies for the high-level data. While our method is designed for robustness to shifts in the exogenous environments, we contaminate the endogenous samples directly. In LAN models, this serves as a direct and computationally efficient proxy, as any additive shift in the exogenous space propagates linearly to an additive shift in the endogenous space. The final performance metric is the expected abstraction error [2], averaged over cross-validation folds, random samples for each noise configuration, and all interventions.

Analysis. We first evaluate all methods under the two extremal settings of our framework: the clean data case, of $\alpha = 0$, and the fully noisy case, of $\alpha = 1$ with a fixed $\tilde{\sigma}$, with results summarized in Table 1. The results consistently demonstrate the value of our robust optimization framework. For clarity of presentation in the main text, we feature the theoretically derived lower bound radius setting $\text{DIROCA}_{\varepsilon_\ell^\star, \varepsilon_h^\star}$ and the consistently well-performing $\text{DIROCA}_{2,2}$ and $\text{DIROCA}_{4,4}$, while full results for more robustness radii are provided in Appendix A.7.

The $\alpha = 0$ setting reveals a nuanced trade-off between standard and robust optimization where the non-robust $\text{GRAD}_{(\tau,\omega)}$ achieves the lowest error across baselines. This highlights the *cost of robustness*: on a clean test dataset, the adversarial training of DIROCA can be overly conservative. Contrarily, under the $\alpha = 1$ case where we fix a $\tilde{\sigma}$ and apply it to all endogenous samples, DIROCA demonstrates superior performance across both settings and datasets, achieving a significantly lower abstraction error than all baselines by yielding more generalizable solutions. This validates our central hypothesis that explicitly accounting for environmental uncertainty during training yields abstractions that are substantially more robust to distributional shifts. To understand the intermediate behavior, we analyze the performance as the fraction of outliers α increases, while holding the noise intensity fixed and vice versa. The results for the SLC and LiLUCAS experiments are shown in Figure 3 and Figure 4 and the ones including all the different DIROCA variants in Appendix A.7 (Figures 10, 11). The plots clearly illustrate that the robust DIROCA variant maintains a lower abstraction error as the data becomes more corrupted or more noise is introduced. At low contamination levels ($\alpha \approx 0$), non-robust methods like $\text{GRAD}_{(\tau,\omega)}$ are outperforming DIROCA variants; however, their error grows sharply as the number of shifted samples or noise increases. $\text{BARY}_{(\tau,\omega)}$ achieves intermediate performance, benefiting from a degree of robustness by aggregating interventions via Wasserstein barycenters, but lacking the explicit adversarial training of DIROCA to guard against worst-case shifts. Notably, the AbsLin baselines underperform overall due to their reliance on observational data and lack of robustness mechanisms. The perfect variant (AbsLin_{p}) fits better under the $\alpha = 0$ case but breaks under distribution shift, while the noisy variant (AbsLin_{n}) generalizes slightly better at the $\alpha = 1$ case thanks to its built-in sparsity regularization. This confirms that their optimization, which is tailored to the clean training data,

[2] In the Gaussian setting, for each dataset X, we first estimate its empirical mean $\hat{\mu}$ and covariance $\hat{\Sigma}$. The Wasserstein distance is then computed between the Gaussian distributions $\mathcal{N}(\hat{\mu}, \hat{\Sigma})$ defined by these estimates.

Table 1: Abstraction error and std under $\alpha = 0$ and $\alpha = 1$ for SLC and LiLUCAS. Top: Gaussian setting with $(\varepsilon_\ell^\star, \varepsilon_h^\star) = (0.11, 0.11)$. Bottom: Empirical setting with $(\varepsilon_\ell^\star, \varepsilon_h^\star) = (0.10, 0.03)$

Method	$\alpha = 0$	$\alpha = 1$	$\alpha = 0$	$\alpha = 1$
	SLC		LiLUCAS	
Gaussian				
$\text{BARY}_{(\tau,\omega)}$	2.63 ± 0.01	8.43 ± 0.01	5.75 ± 0.01	55.11 ± 0.02
$\text{GRAD}_{(\tau,\omega)}$	$\mathbf{1.25 \pm 0.01}$	6.74 ± 0.00	$\mathbf{5.45 \pm 0.02}$	64.73 ± 0.03
$\text{DIROCA}_{\varepsilon_\ell^\star, \varepsilon_h^\star}$	2.67 ± 0.02	11.28 ± 0.01	5.86 ± 0.01	61.86 ± 0.02
$\text{DIROCA}_{2,2}$	1.74 ± 0.01	$\mathbf{2.37 \pm 0.01}$	5.53 ± 0.03	25.79 ± 0.01
$\text{DIROCA}_{4,4}$	1.90 ± 0.01	2.59 ± 0.01	6.81 ± 0.03	$\mathbf{22.99 \pm 0.01}$
Empirical				
$\text{BARY}_{(\tau,\omega)}$	86.46 ± 0.15	650.33 ± 0.69	561.19 ± 3.49	3804.18 ± 4.39
$\text{GRAD}_{(\tau,\omega)}$	$\mathbf{57.04 \pm 0.28}$	392.95 ± 0.30	$\mathbf{309.82 \pm 1.71}$	1290.65 ± 1.52
AbsLin_p	89.29 ± 0.24	678.89 ± 0.71	399.21 ± 1.63	2189.48 ± 2.15
AbsLin_n	91.67 ± 0.29	640.45 ± 0.68	354.31 ± 1.46	1878.96 ± 1.55
$\text{DIROCA}_{\varepsilon_\ell^\star, \varepsilon_h^\star}$	88.39 ± 0.14	674.40 ± 0.74	470.13 ± 3.65	2928.40 ± 3.41
$\text{DIROCA}_{2,2}$	58.86 ± 1.26	$\mathbf{357.93 \pm 1.72}$	315.21 ± 7.93	1102.80 ± 18.15
$\text{DIROCA}_{4,4}$	58.87 ± 1.25	$\mathbf{357.93 \pm 1.72}$	311.19 ± 8.01	$\mathbf{981.61 \pm 18.29}$

fails to generalize when the test distribution is contaminated. The respective analysis for Student-t and Exponential noise can be found in Appendix A.7.

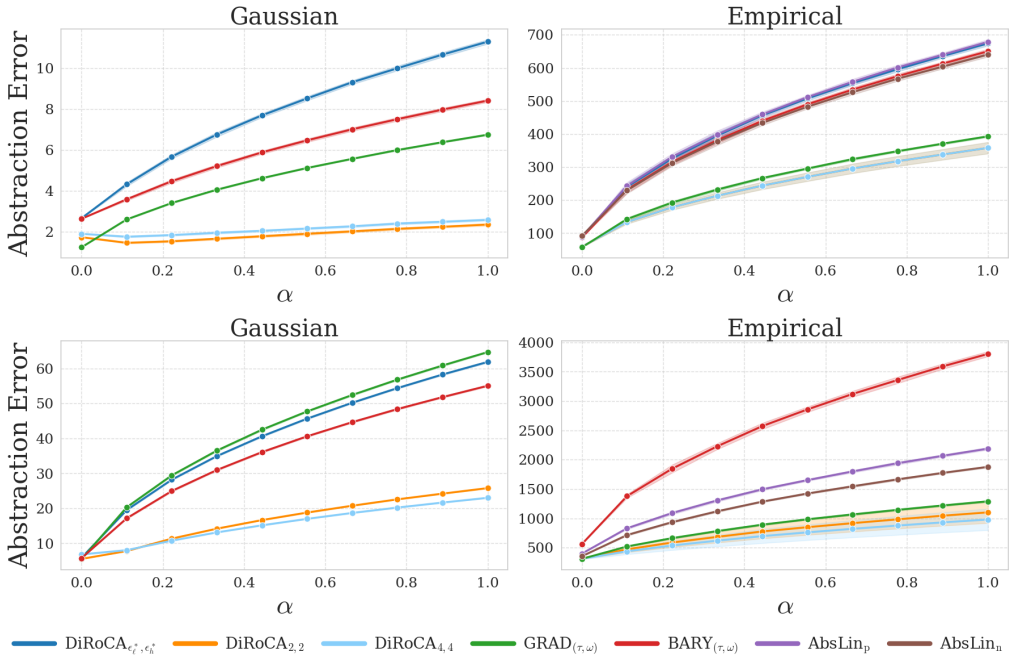


Figure 3: Robustness to outlier fraction (α) on the **SLC** (top) and **LiLUCAS** (bottom) experiments for the Gaussian (left) and Empirical (right) settings. The evaluation is performed at a fixed *Gaussian* noise intensity ($\tilde{\sigma} = 5.0$ for SLC and $\tilde{\sigma} = 10.0$ for LiLUCAS). DIROCA, especially with tuned ambiguity radius, achieves consistently lower abstraction error as the proportion of outlier environments increases, while non-robust methods degrade.

Beyond Environmental Robustness. We also test robustness to violations of core modeling assumptions such as (i) linearity of structural functions ($\mathcal{F}$-misspecification), and (ii) correctness of the intervention mapping ω (ω-misspecification). For $\mathcal{F}$-misspecification, we evaluate our linearly-trained models on new test data generated

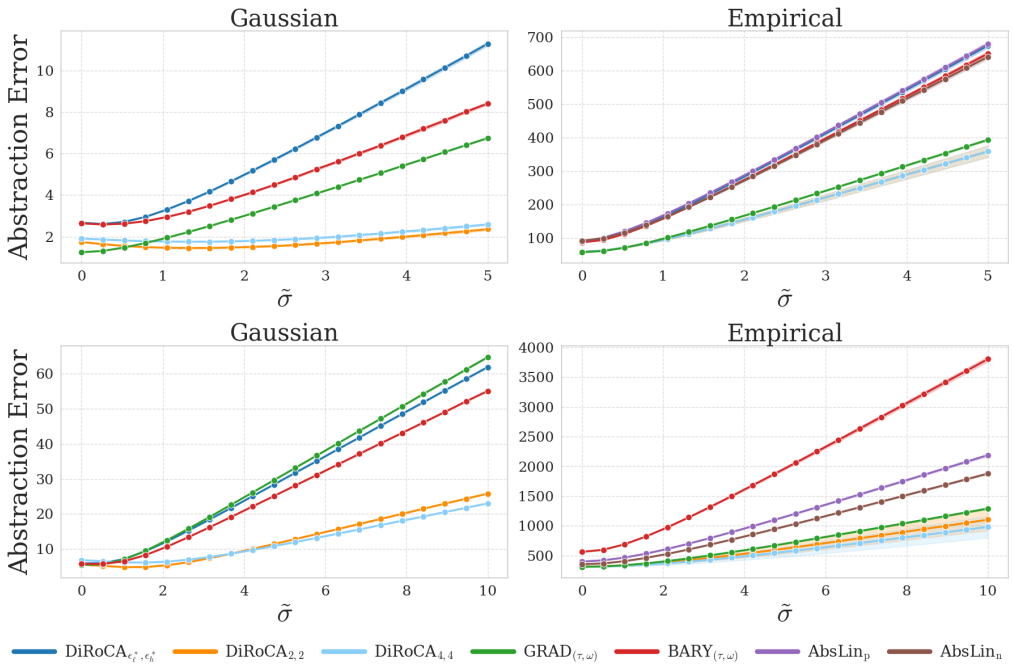


Figure 4: Robustness to *Gaussian* noise intensity ($\tilde{\sigma}$) on the **SLC** (top) and **LiLUCAS** (bottom) experiments for the Gaussian (left) and Empirical (right) settings. The evaluation is performed at a fixed outlier fraction of $\alpha = 1.0$ (fully noisy data). DIROCA achieves consistently lower abstraction error as the proportion of noise intensity increases.

from SCMs with non-linear structural equations. These misspecified SCMs share the same causal graph and environments as those used during training, but the structural equation for each variable is non-linear, controlled by a strength parameter $k \in \mathbb{R}$. For ω-misspecification, we contaminate the ground-truth ω map by reassigning a subset of low-level interventions to different high-level ones of the same complexity (same number of variables), thereby preserving intervention dose while introducing realistic mapping errors. An explanation of both processes can be found in Appendix A.7. In Table 2 we present the results for both the Gaussian and empirical settings of LiLUCAs while the ones of SLC are presented in Appendix A.7. We report the mean and standard deviation of the abstraction error across 100 random trials. For $\mathcal{F}$-misspecification, we report the results for the simplest scenario where $k = 1$ using a sinusoidal (`sin`) structural function. For a more comprehensive view, full robustness curves illustrating the abstraction error as a function of the non-linearity strength k for both `sin` and hyperbolic tangent (`tanh`) functions can be found in Figures 12 and 13 of Appendix A.7. Under both misspecification types, the tuned DIROCA variants consistently outperform all baselines. Specifically, DIROCA$_{2,2}$ attains the lowest error in the Gaussian setting, while DIROCA$_{4,4}$ is the best in the empirical regime. These improvements hold under both misspecification types, demonstrating robustness beyond environmental shifts. This arises from the core min-max design of DIROCA. Optimizing against worst-case distributions within an ambiguity set over the joint environment seems to act as a powerful implicit regularizer. By leveraging the linearity of the SCMs, the algorithm effectively propagates uncertainty from the environment through the system's reduced form to the endogenous level. This process forces the learned abstraction to become resilient to not only environmental shifts but also to violations of the model's own structural assumptions, such as errors in the functional form or the intervention mapping.

6 Conclusion

We introduced a new class of (ρ, ι)-abstractions that ensure interventional consistency across a relevant set of environments. This framework yields robust abstractions that go beyond the limitations of exact abstractions while remaining more tractable than uniform ones. Building on this foundation, we developed the first *robust* Causal Abstraction Learning (CAL) methods, DIROCA, which explicitly account for distributional shifts and misspecification of SCM environments by formulating the problem as a Wasserstein DRO. Our theoretical results extend DRO concentration bounds in the CA setting for the true joint environment under both Gaussian and empirical settings. This provides principled guidance for selecting robustness levels, which also serve as a natural

Table 2: Average abstraction error (mean $\pm\sigma$) under model $\mathcal{F}$-misspecification (left) and intervention mapping ω-misspecification for both the Gaussian with $(\varepsilon_\ell^\star, \varepsilon_h^\star) = (0.11, 0.11)$ and empirical with $(\varepsilon_\ell^\star, \varepsilon_h^\star) = (0.10, 0.03)$ setting of LiLUCAS.

Method	$\mathcal{F}$-misspecification		ω-misspecification	
	Gaussian	Empirical	Gaussian	Empirical
$\text{BARY}_{(\tau,\omega)}$	4.25 ± 0.01	578.80 ± 1.25	5.86 ± 0.03	548.55 ± 2.48
$\text{GRAD}_{(\tau,\omega)}$	5.86 ± 0.03	338.74 ± 1.01	5.63 ± 0.03	305.46 ± 1.31
AbsLin_p	n/a	430.34 ± 1.24	n/a	414.62 ± 1.09
AbsLin_n	n/a	381.32 ± 1.92	n/a	359.45 ± 1.44
$\text{DIROCA}_{\varepsilon_\ell^\star, \varepsilon_h^\star}$	6.01 ± 0.03	500.10 ± 1.44	6.08 ± 0.03	459.06 ± 2.80
$\text{DIROCA}_{2,2}$	$\mathbf{3.99 \pm 0.03}$	333.02 ± 9.86	$\mathbf{5.55 \pm 0.03}$	308.52 ± 8.54
$\text{DIROCA}_{4,4}$	4.63 ± 0.03	$\mathbf{327.42 \pm 9.15}$	6.86 ± 0.02	$\mathbf{304.11 \pm 7.76}$

measure of abstraction strength. Experiments across different problems and prior art demonstrated a consistently lower abstraction error under both distributional shifts and structural misspecification. Notably, we observe that robustness to environmental shifts often induces resilience to broader sources of misspecification, such as imperfect structural assumptions or intervention mappings. The current framework is limited to linear abstractions, depends on reduced-form inversion, and assumes access to both a known intervention map and the true causal DAGs or an accurate estimate via causal discovery. These present challenges that we aim to address in future work.

Acknowledgments

YF acknowledges support by the Onassis Foundation - Scholarship ID: F ZR 063-1/2021-2022. **TD** acknowledges support from a UKRI Turing AI acceleration Fellowship [EP/V02678X/1].

References

Martial Agueh and Guillaume Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011. doi: 10.1137/100805741. URL https://doi.org/10.1137/100805741.

Sander Beckers and Joseph Y Halpern. Abstracting causal models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2678–2685, 2019.

Sander Beckers, Frederick Eberhardt, and Joseph Y Halpern. Approximate causal abstractions. In *Uncertainty in Artificial Intelligence*, pages 606–615. PMLR, 2020.

Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44:375–417, 1991. URL https://api.semanticscholar.org/CorpusID:123428953.

Krzysztof Chalupka, Frederick Eberhardt, and Pietro Perona. Causal feature learning: an overview. *Behaviormetrika*, 44(1):137–164, 2017.

Gabriele D'Acunto, Fabio Massimo Zennaro, Yorgos Felekis, and Paolo Di Lorenzo. Causal abstraction learning based on the semantic embedding principle, 2025. URL https://arxiv.org/abs/2502.00407.

Joel Dyer, Nicholas Bishop, Yorgos Felekis, Fabio Massimo Zennaro, Anisoara Calinescu, Theodoros Damoulas, and Michael Wooldridge. Interventionally consistent surrogates for complex simulation models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 21814–21841. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/26b8e3dc3a21fcd660d80c63b767f324-Paper-Conference.pdf.

Yorgos Felekis, Fabio Massimo Zennaro, Nicola Branchini, and Theodoros Damoulas. Causal optimal transport of abstractions. *Proceedings of CLeaR (Causal Learning and Reasoning) 2024*, 2024.

James Fullwood and Arthur J. Parzygnat. The information loss of a stochastic map. *Entropy*, 23(8), 2021. ISSN 1099-4300. doi: 10.3390/e23081021. URL https://www.mdpi.com/1099-4300/23/8/1021.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, 34:9574–9586, 2021.

Matthias Gelbrich. On a formula for the l2 wasserstein metric between measures on euclidean and hilbert spaces. *Mathematische Nachrichten*, 147(1):185–203, 1990.

Clark R. Givens and R. M. Shortt. A class of wasserstein metrics for probability distributions. *Michigan Mathematical Journal*, 31:231–240, 1984. URL https://api.semanticscholar.org/CorpusID:121338763.

L. Kantorovich. On the transfer of masses (in russian). Doklady Akademii Nauk, August 1942.

Armin Kekić, Bernhard Schölkopf, and Michel Besserve. Targeted reduction of causal models. In *The 40th Conference on Uncertainty in Artificial Intelligence*, 2024. URL https://openreview.net/forum?id=CFHpI53xmb.

Martin Knott and C. S. Smith. On the optimal mapping of distributions. *Journal of Optimization Theory and Applications*, 43:39–49, 1984. URL https://api.semanticscholar.org/CorpusID:120208956.

Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning, 2019. URL https://arxiv.org/abs/1908.08729.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations, 2019. URL https://arxiv.org/abs/1811.12359.

Riccardo Massidda, Sara Magliacane, and Davide Bacciu. Learning causal abstractions of linear structural causal models. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, UAI '24. JMLR.org, 2024.

Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*, pages 666–704, 1781.

Viet Anh Nguyen, Soroosh Shafiee, Damir Filipović, and Daniel Kuhn. Mean-covariance robust risk measurement, 2023. URL https://arxiv.org/abs/2112.09959.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pages 947–1012, 2016.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Eigil F Rischel and Sebastian Weichwald. Compositional abstraction error and a category of causal models. *arXiv preprint arXiv:2103.15758*, 2021.

Eigil Fjeldgren Rischel. The category theory of causal models. Master's thesis, University of Copenhagen, 2020.

Paul K Rubenstein, Sebastian Weichwald, Stephan Bongers, Joris M Mooij, Dominik Janzing, Moritz Grosse-Wentrup, and Bernhard Schölkopf. Causal consistency of structural equation models. In *33rd Conference on Uncertainty in Artificial Intelligence (UAI 2017)*, pages 808–817. Curran Associates, Inc., 2017.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Willem Schooltink and Fabio Massimo Zennaro. Aligning graphical and functional causal abstractions, 2025. URL https://arxiv.org/abs/2412.17080.

Nathan Srebro, Jason Rennie, and Tommi Jaakkola. Maximum-margin matrix factorization. In L. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2004. URL https://proceedings.neurips.cc/paper_files/paper/2004/file/e0688d13958a19e087e123148555e4b4-Paper.pdf.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

Kevin Xia and Elias Bareinboim. Neural causal abstractions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(18):20585–20595, Mar. 2024. doi: 10.1609/aaai.v38i18.30044. URL https://ojs.aaai.org/index.php/AAAI/article/view/30044.

Fabio Massimo Zennaro, Paolo Turrini, and Theo Damoulas. Towards computing an optimal abstraction for structural causal models. In *UAI 2022 Workshop on Causal Representation Learning*, 2022. URL https://openreview.net/forum?id=zGLniPvGsWT.

Fabio Massimo Zennaro, Máté Drávucz, Geanina Apachitei, W. Dhammika Widanage, and Theodoros Damoulas. Jointly learning consistent causal abstractions over multiple interventional distributions. In *2nd Conference on Causal Learning and Reasoning*, 2023. URL https://openreview.net/forum?id=RNs7aMS6zDq.

Fabio Massimo Zennaro, Nicholas George Bishop, Joel Dyer, Yorgos Felekis, Ani Calinescu, Michael J. Wooldridge, and Theodoros Damoulas. Causally abstracted multi-armed bandits. In *The 40th Conference on Uncertainty in Artificial Intelligence*, 2024. URL https://openreview.net/forum?id=Uxrxz4X416.

Pedro C. Álvarez Esteban, E. del Barrio, J.A. Cuesta-Albertos, and C. Matrán. A fixed-point approach to barycenters in wasserstein space. *Journal of Mathematical Analysis and Applications*, 441(2):744–762, 2016. ISSN 0022-247X. doi: https://doi.org/10.1016/j.jmaa.2016.04.045. URL https://www.sciencedirect.com/science/article/pii/S0022247X16300907.

Appendix

A Supplemental material

A.1 Relation to existing CA frameworks.

Fig. 1 provides an overview of how existing causal abstraction (CA) frameworks and learning methods differ in their treatment of environmental variability. The left panel organizes CA frameworks by the subset of the joint environment space $\mathcal{P}(\boldsymbol{\mathcal{U}})$ over which they require interventional consistency.

The (ρ, ι) framework differs from the exact transformation formulation of Rubenstein et al. [2017], and the (R, a, α) of Rischel [2020], both of which implicitly assume a fixed environment $(\mathcal{P}_0(\boldsymbol{\mathcal{U}}))$ without addressing how it constrains or influences the abstraction. As a result, these frameworks may yield ill-defined forms of abstractions when applied beyond that fixed setting, as also noted by Beckers and Halpern [2019], limiting their practical utility. Our notion of (ρ, ι)-abstractions also differs from the *uniform transformations* introduced by Beckers and Halpern [2019], which requires interventional consistency to hold across *all* possible pairs of environments $(\mathcal{P}_\infty(\boldsymbol{\mathcal{U}}))$. By design, uniform transformations are environment-independent: they relate the deterministic components of the SCMs; i.e., their deterministic causal bases without regard to the specific distributions that generate observational data. While uniform abstractions offer a theoretically stronger and philosophically valid notion of abstraction, they are practically infeasible to learn, as they assume access to an unbounded number of environments, many of which may be inaccessible due to physical, data, or knowledge constraints. The (ρ, ι) framework strikes a middle ground: it preserves the formal rigor of exact abstractions while remaining flexible and computationally tractable under real-world distributional limitations by requiring consistency over a constrained relevant subset $\mathcal{P}_m(\boldsymbol{\mathcal{U}})$, with $0 \leqslant m < \infty$. Just as the notion of relevant interventions restricts attention to those that can be meaningfully abstracted and implemented at the high level, we similarly constrain focus to environments that are plausible or meaningful within a given setting. Together, these relevant environments and interventions define the *abstraction context*, which determines the domain over which causal consistency is required.

As discussed, when exact abstractions are not achievable, consistency between models is relaxed to be approximate. However, in the CAL literature, various formulations of abstraction error (Eq. 3) have been proposed. The divergence term $\mathcal{D}_{\mathcal{X}^h}$ has been instantiated using either the Jensen–Shannon divergence [Zennaro et al., 2023, 2024] or the KL divergence [Dyer et al., 2024, Kekić et al., 2024, D'Acunto et al., 2025]. Regarding the aggregation function $\mathfrak{h}$, initial frameworks adopted the maximum over interventions [Beckers et al., 2020, Rischel and Weichwald, 2021], a choice followed by many subsequent CAL works [Zennaro et al., 2023, 2024], while later approaches also considered the expectation over the intervention set [Felekis et al., 2024, Dyer et al., 2024, Kekić et al., 2024]. Note that, especially in policy-making scenarios, some interventions matter more than others due to factors like the likelihood of implementation or cost, reflecting an implicit weighting over the intervention set. While prior works such as [Felekis et al., 2024, Dyer et al., 2024] assume this distribution to be uniform, assigning equal weight to all interventions, it can be adapted to reflect practical priorities or domain-specific preferences.

As for the aggregation function $\mathfrak{g}$, no prior CAL framework has explicitly addressed variability across environments of the two SCMs. The closest suggestion came from Beckers et al. [2020], who, within the setting of uniform abstractions, proposed taking an expectation over exogenous samples of the low-level model. Crucially, they showed the existence of a deterministic map $\tau_{\mathcal{U}}$ that maps low-level exogenous variables into their high-level counterparts. This result relies on a strong assumption, induced by the uniformity property: for any low-level exogenous sample, one can define an environment assigning it probability one, and deterministically map it to the high level (see Appendix, Theorem 3.6 in [Beckers et al., 2020]). Their definition further aggregates over these $\tau_{\mathcal{U}}$ maps via a minimum operator. In contrast, our approach does not assume the existence of a deterministic map across environments. Instead, we treat each environment independently, adopting a data-driven formulation avoiding potentially restrictive assumptions or optimality conditions imposed by a global minimum. That said, since we work with probability measures rather than individual samples, we can always estimate the pushforward of a function, even when its deterministic form is not recoverable. At best, we can approximate a stochastic map derived from the pushforward. In other words, we can always learn a measure-preserving map $\tau_{\mathcal{U}\#} : \rho^{\ell} \to \rho^{h}$, though not necessarily the underlying function $\tau_{\mathcal{U}}$. Of course, computing the pushforward measure comes at a cost: it may be complex, computationally expensive, or difficult to work with. This cost could potentially be interpreted as a form of information loss, an idea further explored in [Fullwood and Parzygnat, 2021].

Regarding the task of CAL, the right panel of Fig. 1 aligns learning methods with their underlying environmental assumptions. Prior CAL approaches operate in the $\epsilon = 0$ regime, implicitly assuming a fixed environment, performing learning under varying settings and assumptions. In contrast, our method introduces a robustness parameter $0 \leqslant \epsilon < \infty$ and explicitly models environmental uncertainty using a Wasserstein ball $\mathbb{B}_{\epsilon}$ centered at the empirical joint environment, enabling generalization to distributional shifts and misspecification within a range controlled by ϵ.

A.2 The Wasserstein Distance

The Wasserstein distance strongly relates to the theory of optimal transport [Villani et al., 2009], [Peyré et al., 2019], which seeks to find the most efficient way of transforming one probability distribution to another. In the classical *Monge formulation* [Monge, 1781], the goal is to find a deterministic map $T : \mathbb{R}^n \to \mathbb{R}^n$ that pushes a source measure μ onto a target measure ν, while minimizing the expected cost of transport, typically taken as the squared Euclidean distance $\|x - T(x)\|_2^2$. However, the Monge problem is highly non-convex and may not admit solutions in general. To address these difficulties, Kantorovich [1942] proposed a relaxation, allowing for *couplings*, these are joint distributions π on $\mathbb{R}^n \times \mathbb{R}^n$ with marginals μ and ν instead of deterministic maps.

The W_2 Wasserstein distance between two probability measures μ and ν on $\mathbb{R}^n$ is then defined as the minimal expected squared cost over all such couplings:

$$W_2(\mu, \nu) := \inf_{\pi \in \Pi(\mu,\nu)} \left(\mathbb{E}_{(X,Y)\sim\pi} \|X - Y\|_2^2\right)^{1/2}, \tag{10}$$

where $\Pi(\mu, \nu)$ denotes the set of all couplings with marginals μ and ν. Intuitively, $W_2(\mu, \nu)$ measures the least amount of effort required to convert μ into ν under a given transport cost.

Wasserstein Distance between Gaussians. When μ and ν are multivariate Gaussians $\mathcal{N}(m_1, \Sigma_1)$ and $\mathcal{N}(m_2, \Sigma_2)$, the 2-Wasserstein distance admits a closed-form expression [Givens and Shortt, 1984]:

$$W_2\left(\mathcal{N}(m_1, \Sigma_1), \mathcal{N}(m_2, \Sigma_2)\right) = \sqrt{\|m_1 - m_2\|_2^2 + \operatorname{Tr}\left(\Sigma_1 + \Sigma_2 - 2\left(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\right)^{1/2}\right)} \tag{11}$$

The squared W_2 is also known as the *Gelbrich distance* [Gelbrich, 1990], and it captures both the displacement between the means and the discrepancy between the covariances. In particular, when the covariance matrices commute, the trace term simplifies to the squared Frobenius norm between their matrix square roots.

Optimal Transport Map between Gaussians. When $\mu = \mathcal{N}(m_1, \Sigma_1)$ and $\nu = \mathcal{N}(m_2, \Sigma_2)$ are two Gaussian measures on $\mathbb{R}^n$, the optimal transport map $T : \mathbb{R}^n \to \mathbb{R}^n$ pushing μ to ν under the W_2 distance is affine and given by [Knott and Smith, 1984]:

$$T(x) = m_2 + A(x - m_1), \tag{12}$$

where the matrix A is symmetric positive definite and is equal to:

$$A = \Sigma_2^{1/2}\left(\Sigma_2^{1/2}\Sigma_1\Sigma_2^{1/2}\right)^{-1/2}\Sigma_2^{1/2}. \tag{13}$$

This map is the gradient of a convex function, and is therefore the unique optimal transport map up to sets of μ-measure zero[3], by Brenier's theorem [Brenier, 1991]. Moreover, it satisfies the relation

$$T_\sharp \mu = \nu, \tag{14}$$

meaning that the pushforward of μ under T is exactly ν.

Wasserstein Barycenters. The Wasserstein barycenter [Agueh and Carlier, 2011] provides a principled way to aggregate a collection of probability distributions into a single representative distribution by minimizing the average squared Wasserstein distance to the given distributions. Given a set of probability measures $\{\nu_j\}_{j=1}^n$ and associated weights $\{\lambda_j\}_{j=1}^n$ with $\lambda_j > 0$ and $\sum_{j=1}^n \lambda_j = 1$, the Wasserstein barycenter $\nu^\star$ is defined as the solution of the following problem:

$$\nu^\star = \arg\min_{\nu} \sum_{j=1}^{n} \lambda_j W_2^2(\nu, \nu_j), \tag{15}$$

where $W_2(\cdot, \cdot)$ denotes the 2-Wasserstein distance between probability measures. In the special case where all ν_j are multivariate Gaussians $\mathcal{N}(\mu_j, \Sigma_j)$, it can be shown that the barycenter $\nu^\star$ is also a Gaussian $\mathcal{N}(\mu^\star, \Sigma^\star)$, where the mean $\mu^\star$ is the weighted average:

$$\mu^\star = \sum_{j=1}^{n} \lambda_j \mu_j \tag{16}$$

and the covariance $\Sigma^\star$ solves the fixed-point equation [Álvarez Esteban et al., 2016]:

$$\Sigma^\star = \sum_{j=1}^{n} \lambda_j \left(\Sigma^{\star 1/2} \Sigma_j \Sigma^{\star 1/2} \right)^{1/2}. \tag{17}$$

Here, the square roots are matrix square roots, and the equation is understood in the positive semidefinite sense.

A.3 Distributionally Robust Optimization (DRO)

Distributionally Robust Optimization (DRO) is a mathematical framework for decision-making under uncertainty, designed to hedge against model misspecification and distributional shifts [Kuhn et al., 2019]. Instead of optimizing performance with respect to a single estimated distribution, DRO considers a set of plausible distributions, called the *ambiguity set*, and seeks solutions that minimize the worst-case expected loss across this set.

Formulation: Let $x \in \mathbb{R}^n$ denote a decision variable, $\xi \in \Xi = \mathbb{R}^m$ a random vector representing uncertain data, and $f(x, \xi) : \mathbb{R}^n \times \Xi \to \mathbb{R}$ a loss function. The standard *Wasserstein DRO* objective is to solve:

$$\inf_{x \in \mathcal{X}} \sup_{\mathbb{Q} \in \mathbb{B}_{\varepsilon,p}(\widehat{\mathbb{P}}_N)} \mathbb{E}_{\xi \sim \mathbb{Q}}[f(x, \xi)], \tag{18}$$

where $\widehat{\mathbb{P}}_N$ is a nominal distribution estimated from data, and $\mathbb{B}_{\varepsilon,p}(\widehat{\mathbb{P}}_N)$ denotes a Wasserstein ball of radius ε centered at $\widehat{\mathbb{P}}_N$:

$$\mathbb{B}_{\varepsilon,p}(\widehat{\mathbb{P}}_N) = \left\{ \mathbb{Q} \in \mathcal{P}(\Xi) : W_p(\mathbb{Q}, \widehat{\mathbb{P}}_N) \leqslant \varepsilon \right\}. \tag{19}$$

The radius ε controls the size of the ambiguity set and reflects the desired level of robustness: a larger ε allows for greater deviations from the nominal distribution (increasing robustness), while a smaller ε yields tighter but potentially less robust decisions. Below, we present two formulations of DRO that are relevant to our work: **(a)** *Elliptical DRO*, which leverages moment-based structural assumptions; and **(b)** *Empirical DRO*, which builds ambiguity sets directly from the empirical data distribution.

Elliptical DRO. Elliptical DRO leverages structural assumptions by modeling the data distribution as belonging to the elliptical family (e.g., Gaussian, Student-t), which is characterized by its mean and covariance. Elliptical distributions are fully characterized by their first two moments, which makes them particularly well-suited for moment-based uncertainty modeling. Specifically, a distribution $\mathbb{Q} = \mathcal{E}_g(\mu, \Sigma)$ is called *elliptical* if it admits a

[3]Formally, uniqueness holds μ-almost everywhere: if two maps both solve the Monge problem, then they agree except possibly on a set of μ-measure zero.

density of the form $f(\xi) = C \det(\Sigma)^{-1} g\left((\xi-\mu)^\top \Sigma^{-1}(\xi-\mu)\right)$, where g is a nonnegative generator function and C is a normalizing constant.

As we saw in the previous section, when $p = 2$, the squared Wasserstein distance between two elliptical distributions admits a closed-form lower bound known as the *Gelbrich distance*:

$$d_G^2((\mu,\Sigma),(\mu',\Sigma')) = \|\mu-\mu'\|^2 + \operatorname{Tr}(\Sigma) + \operatorname{Tr}(\Sigma') - 2\operatorname{Tr}\left((\Sigma^{1/2}\Sigma'\Sigma^{1/2})^{1/2}\right), \tag{20}$$

where (μ,Σ) and (μ',Σ') are the mean-covariance pairs of two distributions.

Assuming elliptical structure, the Wasserstein ambiguity set can be projected onto the space of first and second moments, yielding an *elliptical uncertainty set* [Nguyen et al., 2023]:

$$\mathcal{U}_\varepsilon(\widehat{\mu},\widehat{\Sigma}) = \left\{(\mu,\Sigma) \in \mathbb{R}^m \times \mathbb{S}_+^m : d_G^2\left((\widehat{\mu},\widehat{\Sigma}),(\mu,\Sigma)\right) \leqslant \varepsilon^2\right\}. \tag{21}$$

This construction offers two key advantages: **(i)** interpretability, as uncertainty is captured through shifts in mean and covariance; and **(ii)** computational tractability, since $\mathcal{U}_\varepsilon$ is a convex and compact set. Furthermore, when the nominal distribution $\widehat{\mathbb{P}}_N$ itself is elliptical, the Wasserstein ambiguity set and the elliptical uncertainty set coincide exactly, i.e., $\mathbb{B}_{\varepsilon,p}(\widehat{\mathbb{P}}_N) = \mathcal{U}_\varepsilon(\widehat{\mu},\widehat{\Sigma})$.

Empirical DRO. In contrast, empirical DRO makes no structural assumptions about the underlying data distribution. Instead, the nominal distribution is the empirical measure:

$$\widehat{\mathbb{P}}_N = \frac{1}{N}\sum_{i=1}^N \delta_{\widehat{\xi}_i}, \tag{22}$$

where $\delta_{\widehat{\xi}_i}$ denotes the Dirac measure at the i-th training sample $\widehat{\xi}_i$. A convenient reparameterization restricts attention to ambiguity sets of perturbed empirical distributions of the form:

$$\mathbb{Q}(\Theta) = \frac{1}{N}\sum_{i=1}^N \delta_{\widehat{\xi}_i+\theta_i}, \tag{23}$$

where $\theta_i \in \mathbb{R}^m$ is a displacement vector applied to the i-th sample, and $\Theta = (\theta_1,\ldots,\theta_N) \in \mathbb{R}^{m\times N}$ is the perturbation matrix. The Wasserstein constraint $W_p(\mathbb{Q}(\Theta),\widehat{\mathbb{P}}_N) \leqslant \varepsilon$ is then equivalent to:

$$\frac{1}{N}\sum_{i=1}^N \|\theta_i\|^p \leqslant \varepsilon^p. \tag{24}$$

For $p = 2$, this simplifies to $\|\Theta\|_{\mathrm{F}} \leqslant \varepsilon\sqrt{N}$, where $\|\cdot\|_F$ denotes the Frobenius norm. This formulation offers a clear geometric interpretation: empirical DRO controls the total amount of perturbation allowed on the data samples, measured in aggregate via the Frobenius norm. As a result, optimization under empirical DRO can be interpreted as finding decisions that are robust against small, localized deviations from the observed data points.

A.4 Theorems and Proofs

Proposition 1 (Consistency of Metric-based Abstraction Errors)**.** *Let an abstraction context $(\mathcal{A},\mathcal{I})$ and a distribution q over $\mathcal{I}^\ell$ be given. Suppose $D_{\mathcal{X}^h}$ is a statistical divergence or distance between probability measures satisfying $D_{\mathcal{X}^h}(p,q) = 0 \iff p = q$. If $\mathcal{M}^h$ is a (ρ,ι)-approximate abstraction of $\mathcal{M}^\ell$, then for all $\boldsymbol{\rho} \in \mathcal{A}$ and q-almost surely, $\mathcal{M}^h$ and $\mathcal{M}^\ell$ are (ρ,ι)-interventionally consistent; i.e. Eq.* (2) *holds.*

Proof. We first consider the case where both aggregation functions $\mathfrak{g}$ and $\mathfrak{h}$ are realized as expectations. Specifically, this transforms the environment-intervention error as follows:

$$e_\tau^{\boldsymbol{\rho},\iota}(\mathcal{M}^\ell,\mathcal{M}^h) = \mathbb{E}_{\boldsymbol{\rho}\in\mathcal{A}}\mathbb{E}_{\iota\sim q}\left[\mathcal{D}_{\mathcal{X}^h}\left(\tau_\#(\mathbf{g}^\ell_{\iota\#}(\rho^\ell)),\ \mathbf{g}^h_{\omega(\iota)\#}(\rho^h)\right)\right]. \tag{25}$$

By non-negativity of the divergence $\mathcal{D}_{\mathcal{X}^h}$, we have that for every fixed $\boldsymbol{\rho} \in \mathcal{A}$, the inner expectation is non-negative. Since the total abstraction error is assumed to be zero, it follows that for each $\boldsymbol{\rho} \in \mathcal{A}$,

$$\mathbb{E}_{\iota\sim q}\left[\mathcal{D}_{\mathcal{X}^h}\left(\tau_\#(\mathbf{g}^\ell_{\iota\#}(\rho^\ell)),\ \mathbf{g}^h_{\omega(\iota)\#}(\rho^h)\right)\right] = 0. \tag{26}$$

Taking the expectation of a non-negative random variable implies that

$$\mathcal{D}_{\mathcal{X}^h}\left(\tau_{\#}(\mathbf{g}^{\ell}_{\iota\#}(\rho^{\ell})),\ \mathbf{g}^{h}_{\omega(\iota)\#}(\rho^{h})\right) = 0, \quad q\text{-almost surely} \tag{27}$$

Since $D_{\mathcal{X}^h}(p,q) = 0 \iff p = q$, we conclude that:

$$\tau_{\#}(\mathbf{g}^{\ell}_{\iota\#}(\rho^{\ell})) = \mathbf{g}^{h}_{\omega(\iota)\#}(\rho^{h}) \quad q\text{-almost surely for each fixed } \boldsymbol{\rho} \in \mathcal{A}. \tag{28}$$

If either one or both of the aggregation functions $\mathfrak{g}$ and $\mathfrak{h}$ are realized as an essential supremum, the argument becomes even simpler: since the supremum of a non-negative function is zero if and only if the function itself is zero everywhere, we again conclude that

$$\tau_{\#}\left(\mathbf{g}^{\ell}_{\iota\#}(\rho^{\ell})\right) = \mathbf{g}^{h}_{\omega(\iota)\#}(\rho^{h}) \tag{29}$$

for all $\boldsymbol{\rho} \in \mathcal{A}$ q-almost surely. □

Remark 3. *The proof relies solely on the non-negativity of $D_{\mathcal{X}^h}$ and the fact that $D_{\mathcal{X}^h}(P,Q) = 0$ if and only if $P = Q$. Consequently, the proposition applies to a broad class of divergences and distances, including (but not limited to) the Kullback–Leibler (KL) divergence, the Wasserstein distance W_p for any $p \geqslant 1$, the Total Variation distance, the Hellinger distance, and others.*

Concentration Results for Joint Environments in Causal Abstraction In our causal abstraction setting, each SCM is associated with its own environment: ρ^{ℓ} for the low-level model and ρ^{h} for the high-level model. We assume these environments are independent and define the joint environment as the product measure $\boldsymbol{\rho} = \rho^{\ell} \otimes \rho^{h}$. In practice, however, only empirical samples from each environment are available, leading to an empirical joint distribution $\widehat{\boldsymbol{\rho}} = \widehat{\rho^{\ell}} \otimes \widehat{\rho^{h}}$. It is therefore essential to quantify how well the empirical product distribution approximates the true joint environment. To this end, we establish a concentration result in Wasserstein-2 distance, showing that, with high probability, the true joint environment $\boldsymbol{\rho}$ lies within a 2-Wasserstein product ball centered at $\widehat{\boldsymbol{\rho}}$ for an appropriately chosen radius. This result provides the theoretical foundation for defining a distributionally robust causal abstraction objective over joint uncertainty sets. In particular, the following theorems provide principled guidelines for selecting the radius of the joint ambiguity set, ensuring that it contains the true data-generating process. The results build upon: **(a)** the tensorization property of the Wasserstein distance [Villani et al., 2009]; **(b)** concentration inequalities for elliptical and empirical distributions [Kuhn et al., 2019, Theorems 18 and 21] and the independence assumption between ρ^{ℓ} and ρ^{h}. We first introduce the notion of a light-tailed distribution, which characterizes the tail behavior necessary for applying the concentration results:

Definition A.1 (Light-tailed distribution)**.** A probability distribution $\mathbb{P}$ over $\mathbb{R}^d$ is said to be α*-light-tailed* if there exist constants $\alpha > 0$ and $A > 0$ such that $\mathbb{E}^{\mathbb{P}}\left[\exp\left(\|\xi\|_2^{\alpha}\right)\right] \leqslant A$, where $\xi \sim \mathbb{P}$.

Unlike in [Kuhn et al., 2019], where concentration is studied for a single distribution, our setting involves a product of two independent distributions corresponding to the low- and high-level SCMs. Consequently, our concentration theorems specifically account for this product structure.

Assumption 1. Both ρ^{ℓ} and ρ^{h} are light-tailed distributions.

Theorem 1 (Gaussian $\boldsymbol{\rho}$-Concentration) *Let $\rho^{\ell} \sim \mathcal{N}(\mu_{\ell}, \Sigma_{\ell})$ and $\rho^{h} \sim \mathcal{N}(\mu_{h}, \Sigma_{h})$ under Ass.1, let $\widehat{\rho^{\ell}}$ and $\widehat{\rho^{h}}$ be the empirical distributions from N_{ℓ} and N_{h} i.i.d. samples. Also, let $\boldsymbol{\rho} := \rho^{\ell} \otimes \rho^{h}$ and $\widehat{\boldsymbol{\rho}} := \widehat{\rho^{\ell}} \otimes \widehat{\rho^{h}}$. Then, $\forall\ d \in \{\ell, h\}$, there exist constants $c_d > 0$, depending only on the individual d-environment and confidence levels η_d, such that $\forall \delta \in (0,1]$, with $\delta = 1 - (1-\eta_{\ell})(1-\eta_{h})$:*

$$\mathbb{P}\left[\mathcal{W}_2(\boldsymbol{\rho}, \widehat{\boldsymbol{\rho}}) \leqslant \epsilon\right] \geqslant 1 - \delta, \ \ \forall\ \epsilon \geqslant \sqrt{\left(\frac{\log(c_{\ell}/\eta_{\ell})}{\sqrt{N_{\ell}}}\right)^2 + \left(\frac{\log(c_{h}/\eta_{h})}{\sqrt{N_{h}}}\right)^2}$$

Proof. Our goal is to identify a radius $\epsilon > 0$ such that:

$$\mathbb{P}\left[\mathcal{W}_2(\boldsymbol{\rho}, \widehat{\boldsymbol{\rho}}) \leqslant \epsilon\right] \geqslant 1 - \delta. \tag{30}$$

Let the global event $E_p := \{\mathcal{W}_2(\boldsymbol{\rho}, \widehat{\boldsymbol{\rho}}) \leqslant \epsilon\}$. From [Villani et al., 2009], we know that the squared 2-Wasserstein distance tensorizes for product measures. That is, for $\boldsymbol{\rho} = \rho^{\ell} \otimes \rho^{h}$ and $\widehat{\boldsymbol{\rho}} = \widehat{\rho^{\ell}} \otimes \widehat{\rho^{h}}$, we have:

$$\mathcal{W}_2^2(\boldsymbol{\rho}, \widehat{\boldsymbol{\rho}}) = \mathcal{W}_2^2(\rho^{\ell}, \widehat{\rho^{\ell}}) + \mathcal{W}_2^2(\rho^{h}, \widehat{\rho^{h}}). \tag{31}$$

Therefore, the event E_p is equivalent to the event:

$$\left\{\mathcal{W}_2^2(\rho^\ell, \widehat{\rho^\ell}) + \mathcal{W}_2^2(\rho^h, \widehat{\rho^h}) \leqslant \epsilon^2\right\} \tag{32}$$

Let us now define the per-component local events:

$$E_\ell := \{\mathcal{W}_2(\rho^\ell, \widehat{\rho^\ell}) \leqslant \varepsilon_\ell\} \quad \text{and} \quad E_h := \{\mathcal{W}_2(\rho^h, \widehat{\rho^h}) \leqslant \varepsilon_h\}. \tag{33}$$

If the intersection event $E_\ell \cap E_h$ occurs, then both $\mathcal{W}_2^2(\rho^\ell, \widehat{\rho^\ell}) \leqslant \varepsilon_\ell^2$ and $\mathcal{W}_2^2(\rho^h, \widehat{\rho^h}) \leqslant \varepsilon_h^2$ hold.

Summing these inequalities yields $\mathcal{W}_2^2(\rho^\ell, \widehat{\rho^\ell}) + \mathcal{W}_2^2(\rho^h, \widehat{\rho^h}) \leqslant \varepsilon_\ell^2 + \varepsilon_h^2$. By the tensorization property, this is exactly the condition defining the event E_p and it suffices to select: $\epsilon^2 = \varepsilon_\ell^2 + \varepsilon_h^2$. Thus, the intersection event $E_\ell \cap E_h$ is a subset of the target event E_p:

$$E_\ell \cap E_h \subseteq E_p. \tag{34}$$

Consequently, to establish a lower bound on $\mathbb{P}(E_p)$, it suffices to lower bound the probability of the intersection:

$$\mathbb{P}(E_p) \geqslant \mathbb{P}(E_\ell \cap E_h) \tag{35}$$

However, by construction, the empirical measures $\widehat{\rho^\ell}$ and $\widehat{\rho^h}$ are generated independently from ρ^ℓ and ρ^h. Therefore, the events E_ℓ and E_h are statistically independent. This allows us to factor the probability of the intersection according to the product rule for independent events:

$$\mathbb{P}(E_\ell \cap E_h) = \mathbb{P}(E_\ell)\mathbb{P}(E_h). \tag{36}$$

Next, from Theorem 21 of Kuhn et al. [2019], we have that for any confidence level $\eta_i \in (0, 1]$:

$$\mathbb{P}(E_i) = \mathbb{P}(\mathcal{W}_2(\rho_i, \widehat{\rho}_i) \leqslant \varepsilon_i) \geqslant 1 - \eta_i, \tag{37}$$

whenever $\varepsilon_i \geqslant \tilde{\varepsilon_i} = \frac{\log(c_i/\eta_i)}{\sqrt{N_i}}$, for suitable constants $c_i > 0$ depending on the dimension and tail properties of each distribution ρ_i.

Consequently, by substituting the Eq.s 36 and 37 into the inequality of Eq. 35, we obtain:

$$\mathbb{P}(E_p) \geqslant \mathbb{P}(E_\ell \cap E_h) = \mathbb{P}(E_\ell)\mathbb{P}(E_h) \geqslant (1 - \eta_\ell)(1 - \eta_h). \tag{38}$$

Recalling that $E_p = \{\mathcal{W}_2(\boldsymbol{\rho}, \widehat{\boldsymbol{\rho}}) \leqslant \sqrt{\varepsilon_\ell^2 + \varepsilon_h^2}\}$ and by expressing the desired global confidence $1 - \delta = (1 - \eta_\ell)(1 - \eta_h)$[4], we have shown:

$$\mathbb{P}\left[\mathcal{W}_2(\boldsymbol{\rho}, \widehat{\boldsymbol{\rho}}) \leqslant \epsilon\right] \geqslant (1 - \eta_\ell)(1 - \eta_h) = 1 - \delta. \tag{39}$$

whenever $\epsilon \geqslant \sqrt{\tilde{\varepsilon_\ell}^2 + \tilde{\varepsilon_h}^2}$, completing the proof. □

Theorem 2 (Empirical $\boldsymbol{\rho}$-Concentration) *Let $\widehat{\rho^\ell}$ and $\widehat{\rho^h}$ empirical distributions, under Ass.1, from N_ℓ and N_h i.i.d. samples, with $\boldsymbol{\rho} := \rho^\ell \otimes \rho^h$ and $\widehat{\boldsymbol{\rho}} := \widehat{\rho^\ell} \otimes \widehat{\rho^h}$. Then, for $\forall\, d \in \{\ell, h\}$, there exist constants $c_{d,1}, c_{d,2} > 0$, depending only on the individual d-environment and confidence levels η_d, such that $\forall \delta \in (0, 1]$, with $\delta = 1 - (1 - \eta_\ell)(1 - \eta_h)$, for $N_d(c, \eta) = \log(c_{d,1}/\eta)/c_{d,2}$, if we set:*

$$\tilde{\varepsilon_d} = \left(\frac{\log(c_{d,1}/\eta)}{c_{d,2} N_d}\right)^{\min\{1/d, 1/2\}} \quad \text{if } N_d \geqslant N_d(c, \eta), \quad \text{or} \quad \left(\frac{\log(c_{d,1}/\eta)}{c_{d,2} N_d}\right)^{1/\alpha_d} \quad \text{otherwise,}$$

$$\implies \quad \mathbb{P}\left[\mathcal{W}_2(\boldsymbol{\rho}, \widehat{\boldsymbol{\rho}}) \leqslant \epsilon\right] \geqslant 1 - \delta, \quad \forall\, \epsilon \geqslant \sqrt{\tilde{\varepsilon_\ell}^2 + \tilde{\varepsilon_h}^2}$$

Proof. The core structure of the proof follows exactly that of Theorem 1. Specifically, we define the global event $E_p := \{\mathcal{W}_2(\boldsymbol{\rho}, \widehat{\boldsymbol{\rho}}) \leqslant \epsilon\}$ and local events $E_d := \{\mathcal{W}_2(\rho^d, \widehat{\rho^d}) \leqslant \varepsilon_d\}$, for $d \in \{\ell, h\}$. Using the tensorization property of $\mathcal{W}_2^2$ and the independence of the samples generating $\widehat{\rho^\ell}$ and $\widehat{\rho^h}$, we arrive at the same intermediate conclusion as in the proof of Theorem 1:

$$\mathbb{P}(E_p) \geqslant \mathbb{P}(E_\ell \cap E_h) = \mathbb{P}(E_\ell)\mathbb{P}(E_h), \quad \text{provided that } \epsilon^2 = \varepsilon_\ell^2 + \varepsilon_h^2.$$

[4] A simple choice could be the uniform allocation, given by $\eta_\ell = \eta_d = 1 - (1 - \delta)^{1/2}$.

The key difference from Theorem 1 lies in the specific concentration inequality used to bound the probabilities of the local events $\mathbb{P}(E_d)$. Here, since ρ_d^ℓ are assumed to be general light-tailed distributions, we apply Theorem 18 of Kuhn et al. [2019]. This theorem guarantees that for any $\eta_d \in (0,1]$, $\mathbb{P}(E_d) = \mathbb{P}(\mathcal{W}_2(\rho^d, \widehat{\rho^d}) \leqslant \varepsilon_d) \geqslant 1-\eta_d$, provided that ε_d is chosen to be at least the threshold value $\tilde{\varepsilon_d}$ where:

$$\tilde{\varepsilon_d} := \begin{cases} \left(\dfrac{\log(c_{d,1}/\eta_d)}{c_{d,2}N_d}\right)^{\min\{1/d,1/2\}} & \text{if } N_d \geqslant N_d(c,\eta), \\ \left(\dfrac{\log(c_{d,1}/\eta_d)}{c_{d,2}N_d}\right)^{1/\alpha_d} & \text{otherwise.} \end{cases}$$

To achieve the overall confidence $1-\delta$, we select local confidence levels $\eta_\ell, \eta_h \in (0,1]$ such that $(1-\eta_\ell)(1-\eta_h) = 1-\delta$, and then by substituting the bounds $\mathbb{P}(E_d) \geqslant 1-\eta_d$ into the combined probability inequality, we yield:

$$\mathbb{P}(E_p) \geqslant (1-\eta_\ell)(1-\eta_h) = 1-\delta. \tag{40}$$

Thus, the result $\mathbb{P}\left[\mathcal{W}_2(\boldsymbol{\rho}, \widehat{\boldsymbol{\rho}}) \leqslant \epsilon\right] \geqslant 1-\delta$ is established once again, whenever $\epsilon \geqslant \sqrt{\tilde{\varepsilon_\ell}^2 + \tilde{\varepsilon_h}^2}$. □

Remark 4. *These results justify the use of joint Wasserstein ambiguity sets centered at the empirical product environment. They ensure that, with high probability, the true environment lies within a ball of radius ϵ, allowing us to formulate a robust abstraction objective that is consistent with finite-sample uncertainty. The construction is especially well-suited for our setting, where only a single environment is available from each SCM.*

A.5 The Joint Ambiguity Set

Here, we provide a visual and formal illustration of the construction of the *joint ambiguity set* used in our framework, as described in the main text. Our goal is to introduce distributional robustness at the level of environments by following the Distributionally Robust Optimization (DRO) paradigm. Since we assume access to a single observed environment from each SCM (low-level and high-level), we model distributional uncertainty using a 2-Wasserstein product ball centered at the empirical joint environment $\widehat{\boldsymbol{\rho}}$ that is formally defined as the joint ambiguity set:

$$\mathbb{B}_{\epsilon,2}(\widehat{\boldsymbol{\rho}}) := \mathbb{B}_{\varepsilon_\ell,2}(\widehat{\rho^\ell}) \times \mathbb{B}_{\varepsilon_h,2}(\widehat{\rho^h}), \tag{41}$$

where for each domain $d \in \{|\mathcal{M}^\ell|, |\mathcal{M}^h|\}$, the marginal ambiguity set $\mathbb{B}_{\varepsilon_d,2}(\widehat{\rho^d})$ is a 2-Wasserstein ball defined as

$$\mathbb{B}_{\varepsilon_d,2}(\widehat{\rho^d}) = \left\{\rho \in \mathcal{P}(\mathcal{U}^d) : W_2(\rho, \widehat{\rho^d}) \leqslant \varepsilon_d\right\}. \tag{42}$$

Each radius ε_d controls the robustness level for the corresponding SCM, with larger values providing robustness to broader shifts but may lead to more conservative solutions.

Construction Steps. Since we do not observe exogenous environments directly, we reconstruct them via abduction. Specifically, given samples from the endogenous variables, we invert the reduced form of the SCMs to recover approximate exogenous samples:

$$(\mathbf{g}_\#^d)^{-1}(\widehat{\mathcal{X}^d}) \approx \widehat{\mathcal{U}^d}, \quad \text{for } d \in \{|\mathcal{M}^\ell|, |\mathcal{M}^h|\}. \tag{43}$$

In Linear Additive Noise (LAN) models with known structural coefficients, this inversion is exact. Otherwise, if the structural functions must be estimated, the quality of the abduction depends on the estimation accuracy of Linear Regression. After performing abduction, we form the empirical exogenous distributions $\widehat{\mathcal{U}^\ell}$ and $\widehat{\mathcal{U}^h}$ via empirical measures:

$$\widehat{\mathcal{U}^\ell} = \frac{1}{N}\sum_{i=1}^{N} \delta_{\widehat{\mathbf{u}_i^L}}, \quad \widehat{\mathcal{U}^h} = \frac{1}{M}\sum_{j=1}^{M} \delta_{\widehat{\mathbf{u}_j^H}}, \tag{44}$$

where $\widehat{\mathbf{u}_i^L}$ and $\widehat{\mathbf{u}_j^H}$ denote the reconstructed exogenous samples at the low and high levels, respectively. Next, we form the nominal joint distribution $\widehat{\boldsymbol{\rho}} := \widehat{\rho^\ell} \otimes \widehat{\rho^h}$ by combining the empirical exogenous distributions. Finally, we build the joint ambiguity set $\mathbb{B}_{\epsilon,2}(\widehat{\boldsymbol{\rho}})$ by taking the product of two marginal Wasserstein balls. The ambiguity set $\mathbb{B}_{\epsilon,2}(\widehat{\boldsymbol{\rho}})$ allows us to find the worst-case joint environment $\boldsymbol{\rho}^\star = \rho^{\ell\star} \otimes \rho^{h\star}$ that maximizes the abstraction error within the specified robustness radii. This captures the most adversarial distributional shift consistent with the observed data and estimated environments. The radii ε_ℓ and ε_h quantify the maximum deviation we aim to protect against at each abstraction level. The full construction process, including the abduction, nominal estimation, and radius selection steps, is illustrated in Fig. 5 for the case where low- and high-level environments are sampled independently.

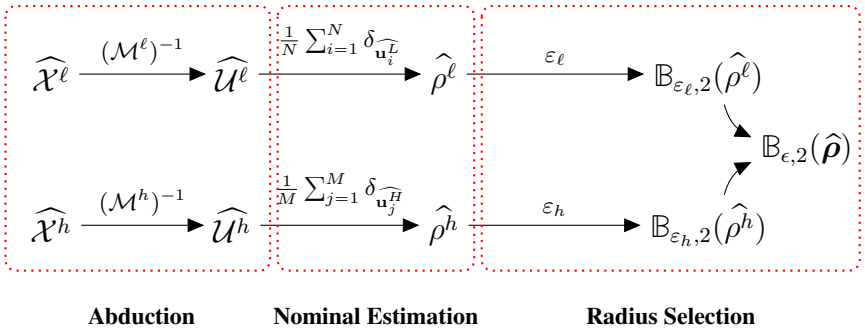


Figure 5: Construction of the joint ambiguity set $\mathbb{B}_{\epsilon,2}(\widehat{\boldsymbol{\rho}})$ under the assumption of independently sampled environments. The process involves abduction of exogenous variables, nominal empirical distribution estimation, and selection of robustness radii.

A.6 Datasets

Simple Lung Cancer (SLC) Dataset The Simple Lung Cancer (SLC) dataset models the causal relationships between three continuous variables: S denotes *Smoking*, T denotes *Tar deposition* in the lungs, and C denotes *Cancer* development. In the low-level representation, smoking causes tar accumulation, which in turn causes cancer. The high-level abstraction removes the intermediate variable T, resulting in a direct causal link between smoking and cancer. The experiment includes a set of 6 distinct low-level and 3 high-level binary interventions. The corresponding low-level and high-level causal graphs are shown in Fig. 6.

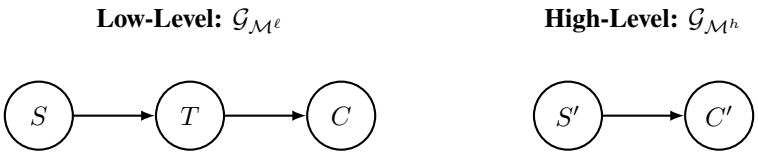


Figure 6: Low-level ($\mathcal{G}_{\mathcal{M}^\ell}$) and high-level ($\mathcal{G}_{\mathcal{M}^h}$) causal graphs for the Simple Lung Cancer (SLC) dataset. In the high-level abstraction, the intermediate variable T (tar deposition) is omitted, resulting in a direct causal link between smoking and cancer.

Linearized LUCAS (LiLUCAS) Dataset The Linearized LUCAS (LiLUCAS) dataset is a simplified, linearized version of the LUng CAncer dataset (LUCAS)[5], originally designed to simulate realistic causal relationships in lung cancer diagnosis. The low-level model involves several continuous variables: SM denotes *Smoking*, GE denotes *Genetics*, LC denotes *Lung Cancer*, AL denotes *Allergy*, CO denotes *Coughing*, and FA denotes *Fatigue*. In the high-level abstraction, groups of variables are clustered into broader concepts: EN$'$ (Environment), GE$'$ (Genetics), and LC$'$ (Lung Cancer). It involves a set of 20 distinct low-level and 11 high-level binary interventions. The underlying low-level and high-level causal graphs are shown in Fig. 7.

A.7 Additional experimental results and settings

We evaluate our framework across a diverse set of experimental settings of different representations of the environment (Gaussian vs. empirical) and datasets. Fig. 8 provides an overview of the datasets, methods, and evaluation metrics used in each setting.

A.7.1 Misspecifications and settings

Robustness to Distributional Shifts. We evaluate robustness using a unified and flexible framework based on a Huber-style contamination model, applied to the held-out test data from each of the k cross-validation folds. The methodology consists of data contamination, error computation, and a multi-level aggregation protocol.

1. Data Contamination. For a given clean interventional test set, $X_{\text{test},\iota} \in \mathbb{R}^{n_{\text{test}} \times d}$, we first generate a noisy outlier set, $\tilde{X}_\iota$, by applying an additive stochastic shift. This is achieved by generating a noise matrix

[5] http://www.causality.inf.ethz.ch/data/LUCAS.html

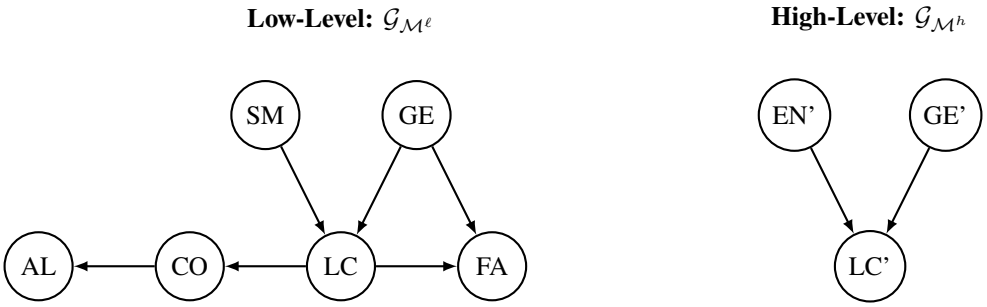


Figure 7: Low-level ($\mathcal{G}_{\mathcal{M}^\ell}$) and high-level ($\mathcal{G}_{\mathcal{M}^h}$) causal graphs for the Linearized LUCAS (LiLUCAS) dataset. The abstraction groups variables related to environment, genetics, and disease outcomes, resulting in a simplified causal structure.

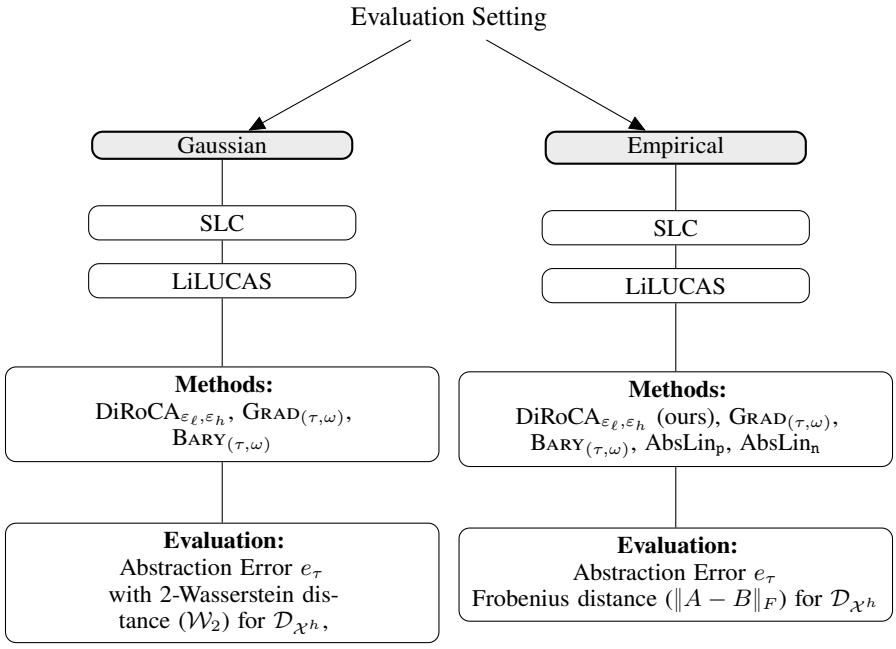


Figure 8: Roadmap of experimental settings.

$\mathbf{N} \in \mathbb{R}^{n_{\text{test}} \times d}$, where each row is an independent sample from a specified distribution. While the main text focuses on Gaussian noise ($\mathbf{n}_i \sim \mathcal{N}(0, \sigma^2_{\text{noise}} \cdot \mathbf{I})$), our framework also supports other distributions, including heavy-tailed Student-t and skewed Exponential. The final contaminated test set, $\bar{X}_\iota$, is then formed as a convex combination of the clean and noisy data, governed by the contamination fraction $\alpha \in [0, 1]$:

$$\bar{X}_\iota = (1-\alpha)X_{\text{test},\iota} + \alpha\tilde{X}_\iota. \tag{45}$$

This process is controlled by two parameters: the fraction α (prevalence) and the noise scale σ_{noise} (magnitude).

2. Error Computation. For a single contaminated test set and a given abstraction map T, the abstraction error is computed as the average distance between the transformed low-level and high-level samples, averaged across all interventions $\iota \in \mathcal{I}_\ell$:

$$\mathcal{E}(\tau, \alpha, \sigma_{\text{noise}}) = \mathbb{E}_{\iota\sim q}\left[\mathcal{D}\left(\mathrm{T}\,\bar{X}^\ell_\iota, \bar{X}^h_{\omega(\iota)}\right)\right]. \tag{46}$$

3. Experimental Protocol. To ensure statistically robust results, we perform a multi-level evaluation. The full experiment is a nested procedure that iterates through a grid of contamination parameters (α and $\tilde{\sigma}$).

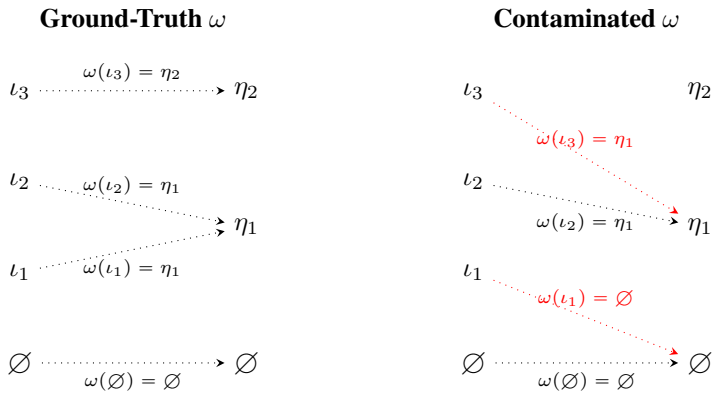


Figure 9: Illustration of intervention map (ω) contamination. Left: the ground-truth mapping between low-level interventions (ι) and high-level interventions (η). Right: a contaminated version where interventions have been re-assigned to incorrect counterparts of similar complexity (e.g., $\omega(\iota_1) = \eta_2$ instead of η_1) shown in red.

For each point on this grid, our protocol yields a total of $m \times k$ individual error measurements (one for each of the m random samples on each of the k data splits). The final "Robustness Curves" visualize the mean error, averaged over both samples and folds, as a function of the contamination parameters.

Robustness to Model Misspecification. Beyond robustness to environmental shifts, we evaluate our method's resilience to violations of core modeling assumptions. We challenge two critical assumptions: (a) the linearity of the structural functions ($\mathcal{F}$-misspecification), and (b) the correctness of the intervention mapping (ω-misspecification).

- **$\mathcal{F}$-misspecification.** To test robustness against violations of the linearity assumption, we evaluate our linearly-trained models on test data generated from fundamentally non-linear SCMs. Crucially, these new SCMs share the same causal graph, intervention sets, and underlying environments as those used in training. However, we define a new structural non-linear equations f_{NL}, controlled by a strength parameter $k \in \mathbb{R}$:

$$X_j = k \cdot f_{\text{NL}}(\text{Pa}(X_j)) + U_j. \tag{47}$$

  For each strength level k, we simulate a brand new, ground-truth non-linear test set from this SCM and compute the abstraction error. The main text reports the results for $k = 1$ using a sinusoidal function. Here we also provide full robustness curves showing the error as a function of $k \in [0, 100]$ for both `sin` and `tanh` non-linearities in Figures 12 and 13 respectively.
- **ω-misspecification.** To test robustness to an incorrect intervention mapping, we contaminate the ground-truth ω map using a semantic corruption strategy. For a given number of misalignments, we randomly select a subset of low-level interventions and re-assign each one to an incorrect high-level counterpart of similar complexity (i.e., one that intervenes on a similar number of variables), with a fallback to the nearest complexity neighbor if an exact match is not available within a specified delta. An illustration of this semantic ω-misspecification process is provided in Figure 9.

Implementation and Optimization Details. All models were trained on a MacBook Air (13-inch, M1, 2020) equipped with an Apple M1 chip and 16 GB of unified memory, running macOS Monterey version 12.3. Performance is evaluated using a k-fold cross-validation scheme (with $k = 5$) on the 10,000 generated samples for each intervention in both *LiLUCAS* and *SLC*. In DIROCA each iteration comprises 5 minimization and 2 maximization steps. Finally, we monitored convergence by tracking the absolute change in the objective value across successive epochs, and the optimization is terminated once this change falls below a fixed threshold of 10^{-4}, ensuring numerical stability while avoiding unnecessary computation.

A.7.2 Results

Below, we present the full experimental results as outlined in the experiments roadmap (Fig. 8). All bolded values in the tables indicate statistically significant differences, determined using paired t-tests with a significance threshold of $p < 0.05$.

Table 3: Abstraction error and std under $\alpha = 0$ and $\alpha = 1$ for SLC and LiLUCAS. Top: Gaussian setting with $(\varepsilon_\ell^\star, \varepsilon_h^\star) = (0.11, 0.11)$. Bottom: Empirical setting with $(\varepsilon_\ell^\star, \varepsilon_h^\star) = (0.10, 0.03)$

Method	$\alpha = 0$	$\alpha = 1$	$\alpha = 0$	$\alpha = 1$
	SLC		LiLUCAS	
	Gaussian			
$\text{BARY}_{(\tau,\omega)}$	2.63 ± 0.01	8.43 ± 0.01	5.75 ± 0.01	55.11 ± 0.02
$\text{GRAD}_{(\tau,\omega)}$	$\mathbf{1.25 \pm 0.01}$	6.74 ± 0.00	$\mathbf{5.45 \pm 0.02}$	64.73 ± 0.03
$\text{DIROCA}_{\varepsilon_\ell^\star, \varepsilon_h^\star}$	2.67 ± 0.02	11.28 ± 0.01	5.86 ± 0.01	61.86 ± 0.02
$\text{DIROCA}_{1,1}$	2.01 ± 0.01	4.14 ± 0.01	6.97 ± 0.03	48.05 ± 0.02
$\text{DIROCA}_{2,2}$	1.74 ± 0.01	$\mathbf{2.37 \pm 0.01}$	5.53 ± 0.03	25.79 ± 0.01
$\text{DIROCA}_{4,4}$	1.90 ± 0.01	2.59 ± 0.01	6.81 ± 0.03	$\mathbf{22.99 \pm 0.01}$
$\text{DIROCA}_{8,8}$	1.90 ± 0.01	2.59 ± 0.01	6.81 ± 0.03	$\mathbf{23.00 \pm 0.01}$
	Empirical			
$\text{BARY}_{(\tau,\omega)}$	86.46 ± 0.15	650.33 ± 0.69	561.19 ± 3.49	3804.18 ± 4.39
$\text{GRAD}_{(\tau,\omega)}$	$\mathbf{57.04 \pm 0.28}$	392.95 ± 0.30	$\mathbf{309.82 \pm 1.71}$	1290.65 ± 1.52
AbsLin_p	89.29 ± 0.24	678.89 ± 0.71	399.21 ± 1.63	2189.48 ± 2.15
AbsLin_n	91.67 ± 0.29	640.45 ± 0.68	354.31 ± 1.46	1878.96 ± 1.55
$\text{DIROCA}_{\varepsilon_\ell^\star, \varepsilon_h^\star}$	88.39 ± 0.14	674.40 ± 0.74	470.13 ± 3.65	2928.40 ± 3.41
$\text{DIROCA}_{1,1}$	59.08 ± 0.33	388.73 ± 0.31	332.96 ± 2.98	1385.36 ± 1.58
$\text{DIROCA}_{2,2}$	58.86 ± 1.26	$\mathbf{357.93 \pm 1.72}$	315.21 ± 7.93	1102.80 ± 18.15
$\text{DIROCA}_{4,4}$	58.87 ± 1.25	$\mathbf{357.93 \pm 1.72}$	311.19 ± 8.01	$\mathbf{981.61 \pm 18.29}$
$\text{DIROCA}_{8,8}$	58.89 ± 1.21	358.20 ± 1.67	311.43 ± 8.18	987.53 ± 17.81

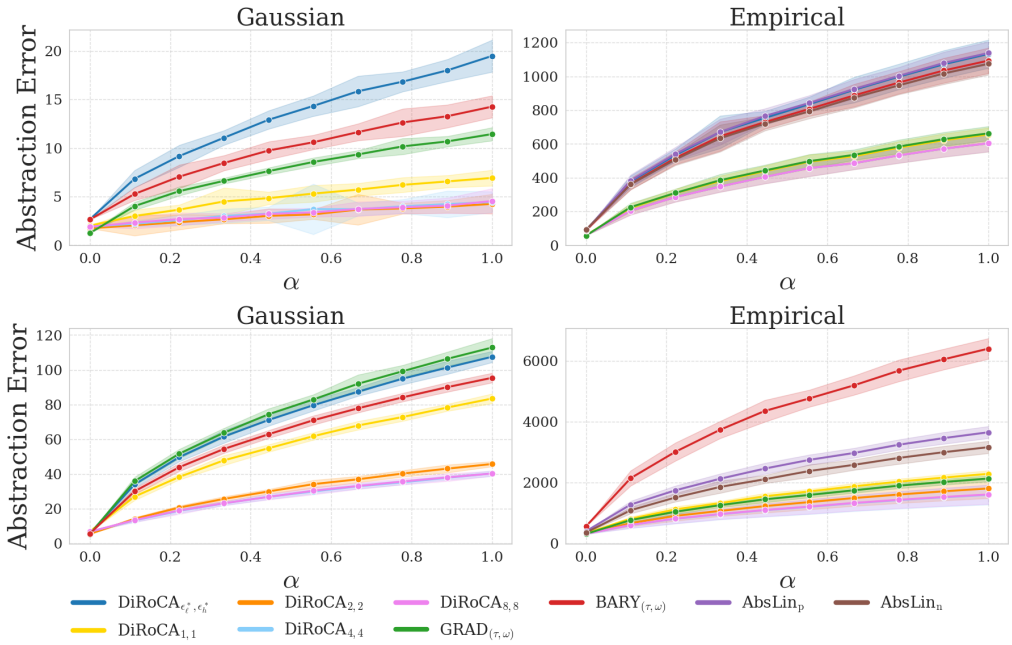


Figure 10: Robustness to outlier fraction (α) on the **SLC** (top) and **LiLUCAS** (bottom) experiments for the Gaussian (left) and Empirical (right) settings. The evaluation is performed at a fixed *Gaussian* noise intensity ($\tilde{\sigma} = 5.0$ for SLC and $\tilde{\sigma} = 10.0$ for LiLUCAS)

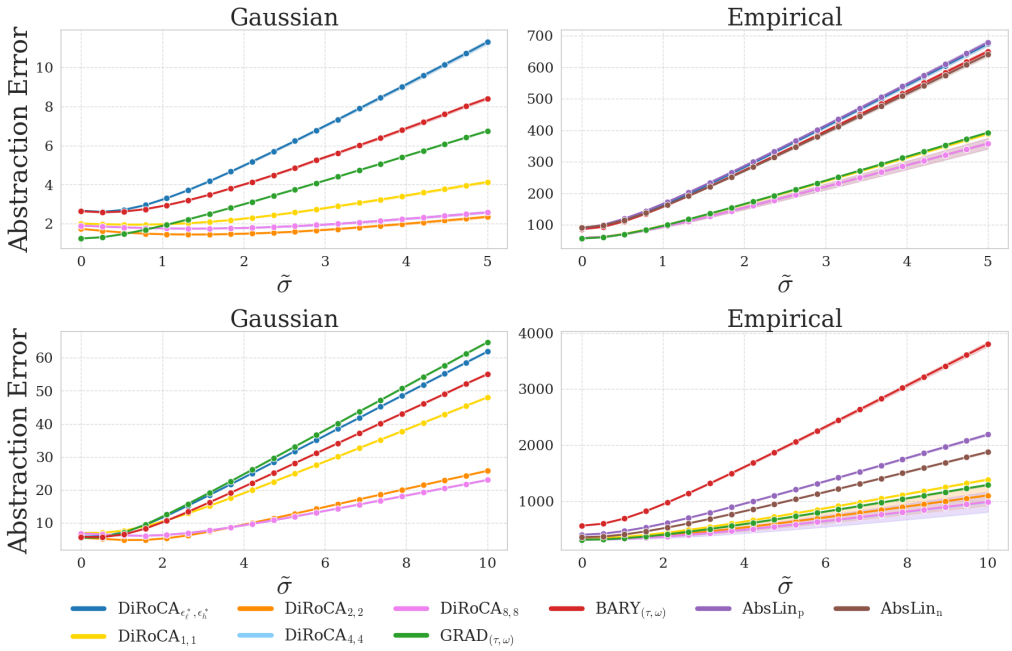


Figure 11: Robustness to *Gaussian* noise intensity ($\tilde{\sigma}$) on the **SLC** (top) and **LiLUCAS** (bottom) experiments for the Gaussian (left) and Empirical (right) settings. The evaluation is performed at a fixed outlier fraction of $\alpha = 1.0$ (fully noisy data).

Table 4: Average abstraction error (mean $\pm\sigma$) under *model* $\mathcal{F}$-misspecification with `sin` nonlinearity for the Gaussian setting with $(\varepsilon_\ell^\star, \varepsilon_h^\star) = (0.11, 0.11)$ and the empirical setting with $(\varepsilon_\ell^\star, \varepsilon_h^\star) = (0.10, 0.03)$.

Method	SLC		LiLUCAS	
	Gaussian	Empirical	Gaussian	Empirical
BARY$_{(\tau,\omega)}$	1.10 $\pm$ 0.02	185.84 $\pm$ 1.50	4.25 $\pm$ 0.01	578.80 $\pm$ 1.25
GRAD$_{(\tau,\omega)}$	2.96 $\pm$ 0.01	138.33 $\pm$ 0.96	5.86 $\pm$ 0.03	338.74 $\pm$ 1.01
AbsLin$_{\text{p}}$	n/a	198.39 $\pm$ 1.70	n/a	430.34 $\pm$ 1.24
AbsLin$_{\text{n}}$	n/a	190.58 $\pm$ 1.67	n/a	381.32 $\pm$ 1.92
DIROCA$_{\varepsilon_\ell^\star,\varepsilon_h^\star}$	**0.40 $\pm$ 0.02**	191.68 $\pm$ 1.54	6.01 $\pm$ 0.03	500.10 $\pm$ 1.44
DIROCA$_{1,1}$	1.79 $\pm$ 0.02	137.37 $\pm$ 0.92	5.98 $\pm$ 0.03	353.61 $\pm$ 1.83
DIROCA$_{2,2}$	1.43 $\pm$ 0.01	135.34 $\pm$ 0.51	**3.99 $\pm$ 0.03**	333.02 $\pm$ 9.86
DIROCA$_{4,4}$	2.19 $\pm$ 0.02	135.34 $\pm$ 0.50	4.63 $\pm$ 0.03	**327.42 $\pm$ 9.15**
DIROCA$_{8,8}$	2.19 $\pm$ 0.02	**135.32 $\pm$ 0.52**	4.63 $\pm$ 0.03	327.64 $\pm$ 9.00

Table 5: Average abstraction error (mean $\pm\sigma$) under *model* $\mathcal{F}$-misspecification with `tanh` nonlinearity for the Gaussian setting with $(\varepsilon_\ell^\star, \varepsilon_h^\star) = (0.11, 0.11)$ and the empirical setting with $(\varepsilon_\ell^\star, \varepsilon_h^\star) = (0.10, 0.03)$.

Method	SLC		LiLUCAS	
	Gaussian	Empirical	Gaussian	Empirical
BARY$_{(\tau,\omega)}$	1.22 $\pm$ 0.01	186.94 $\pm$ 1.39	**3.85 $\pm$ 0.01**	564.09 $\pm$ 1.81
GRAD$_{(\tau,\omega)}$	2.99 $\pm$ 0.01	140.07 $\pm$ 0.87	5.42 $\pm$ 0.04	338.69 $\pm$ 0.93
AbsLin$_{\text{p}}$	n/a	195.23 $\pm$ 1.62	n/a	444.13 $\pm$ 1.27
AbsLin$_{\text{n}}$	n/a	191.79 $\pm$ 1.57	n/a	386.72 $\pm$ 2.03
DIROCA$_{\varepsilon_\ell^\star,\varepsilon_h^\star}$	**0.55 $\pm$ 0.01**	192.73 $\pm$ 1.44	5.61 $\pm$ 0.04	486.29 $\pm$ 1.89
DIROCA$_{1,1}$	2.05 $\pm$ 0.02	138.89 $\pm$ 0.84	5.86 $\pm$ 0.04	352.50 $\pm$ 1.84
DIROCA$_{2,2}$	1.57 $\pm$ 0.01	136.98 $\pm$ 0.42	4.08 $\pm$ 0.03	333.73 $\pm$ 8.63
DIROCA$_{4,4}$	2.32 $\pm$ 0.02	136.99 $\pm$ 0.42	4.96 $\pm$ 0.03	**328.82 $\pm$ 8.26**
DIROCA$_{8,8}$	2.32 $\pm$ 0.02	**136.92 $\pm$ 0.45**	4.96 $\pm$ 0.03	329.04 $\pm$ 8.12

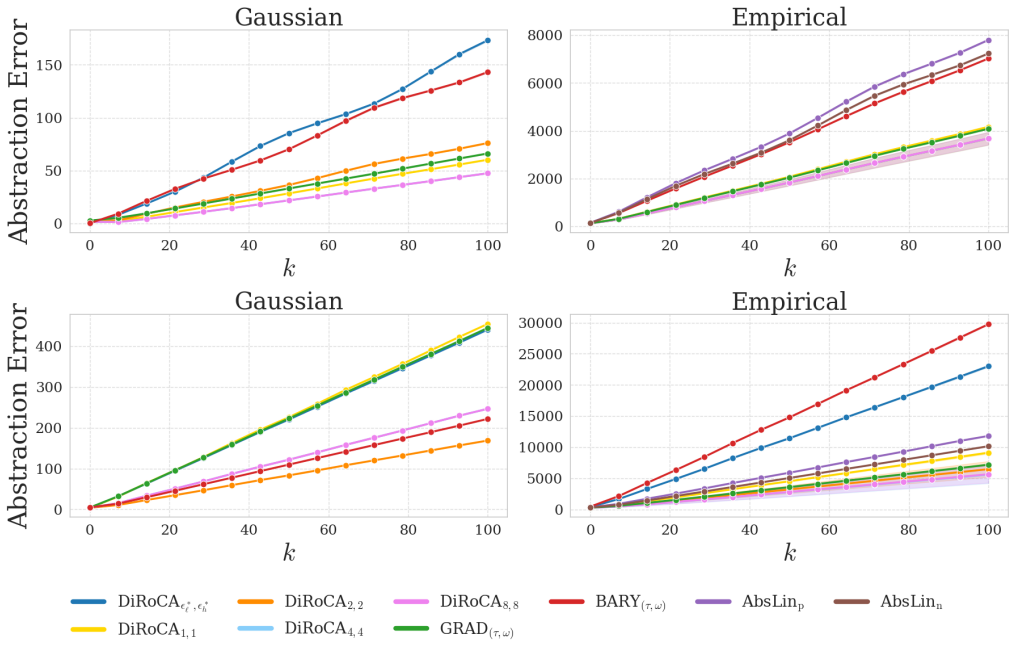


Figure 12: Abstraction error comparison as a function of the nonlinearity strength parameter k, which controls the strength of the nonlinear component (`sin`) in the data generation process. The figure shows results for both Gaussian (left) and empirical (right) data distributions in the SLC (top) and LiLUCAS (bottom) settings.

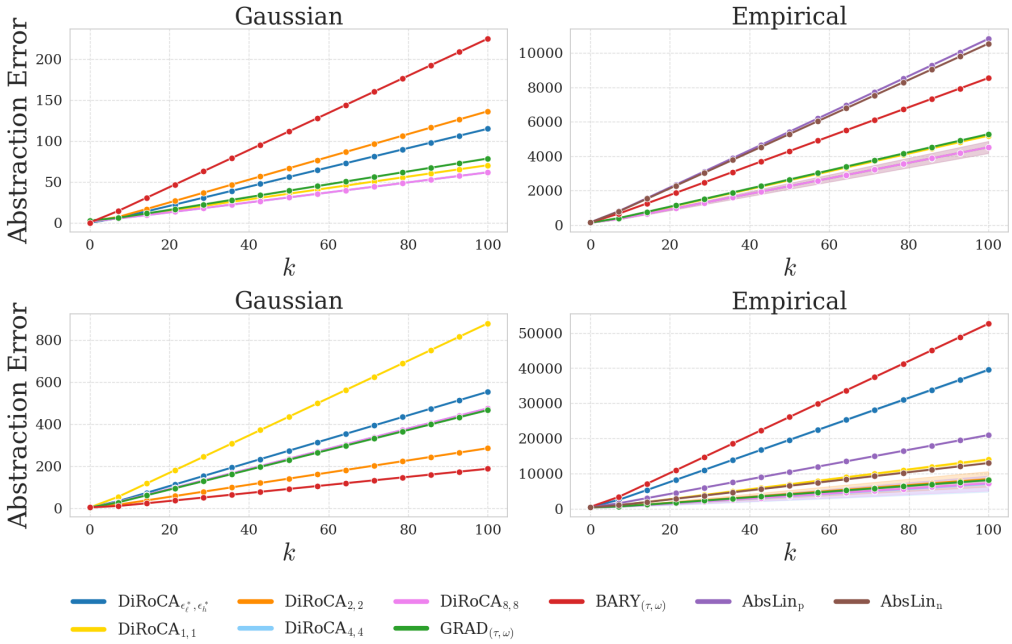


Figure 13: Abstraction error comparison as a function of the nonlinearity strength parameter k, which controls the strength of the nonlinear component (`tanh`) in the data generation process. The figure shows results for both Gaussian (left) and empirical (right) data distributions in the SLC (top) and LiLUCAS (bottom) settings.

Table 6: Average abstraction error (mean $\pm\sigma$) under *intervention mapping* ω-misspecification for the Gaussian setting with $(\varepsilon_\ell^\star, \varepsilon_h^\star) = (0.11, 0.11)$ and the empirical setting with $(\varepsilon_\ell^\star, \varepsilon_h^\star) = (0.10, 0.03)$.

Method	SLC		LiLUCAS	
	Gaussian	Empirical	Gaussian	Empirical
$\text{BARY}_{(\tau,\omega)}$	2.77 ± 0.00	100.98 ± 0.24	5.86 ± 0.03	548.55 ± 2.48
$\text{GRAD}_{(\tau,\omega)}$	$\mathbf{0.71} \pm \mathbf{0.00}$	50.39 ± 0.15	5.63 ± 0.03	305.46 ± 1.31
AbsLin_p	n/a	111.24 ± 0.17	n/a	414.62 ± 1.09
AbsLin_n	n/a	105.50 ± 0.14	n/a	359.45 ± 1.44
$\text{DIROCA}_{\varepsilon_\ell^\star, \varepsilon_h^\star}$	2.57 ± 0.02	105.03 ± 0.24	6.08 ± 0.03	459.06 ± 2.80
$\text{DIROCA}_{1,1}$	1.54 ± 0.00	46.21 ± 0.14	7.16 ± 0.03	326.56 ± 2.59
$\text{DIROCA}_{2,2}$	1.48 ± 0.01	$\mathbf{39.16} \pm \mathbf{2.75}$	$\mathbf{5.55} \pm \mathbf{0.03}$	308.52 ± 8.54
$\text{DIROCA}_{4,4}$	1.43 ± 0.00	$\mathbf{39.16} \pm \mathbf{2.75}$	6.86 ± 0.02	$\mathbf{304.11} \pm \mathbf{7.76}$
$\text{DIROCA}_{8,8}$	1.43 ± 0.00	39.28 ± 2.51	6.86 ± 0.02	304.32 ± 7.83

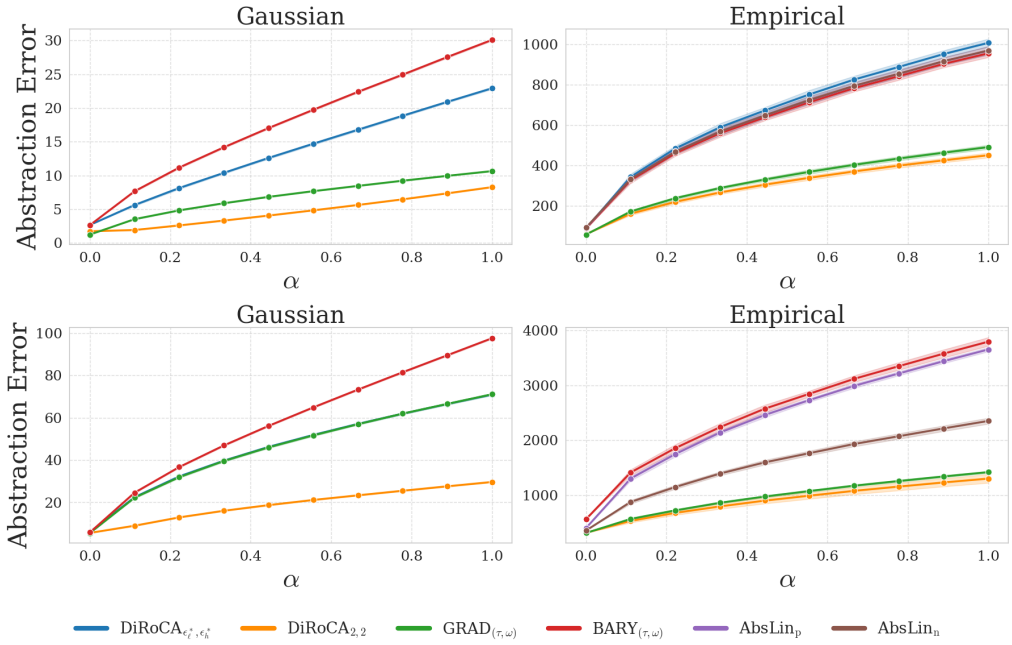


Figure 14: Robustness to outlier fraction (α) on the **SLC** (top) and **LiLUCAS** (bottom) experiments for the Gaussian (left) and Empirical (right) settings. The evaluation is performed at a fixed *Exponential* noise intensity ($\tilde{\sigma} = 5.0$ for SLC and $\tilde{\sigma} = 10.0$ for LiLUCAS)

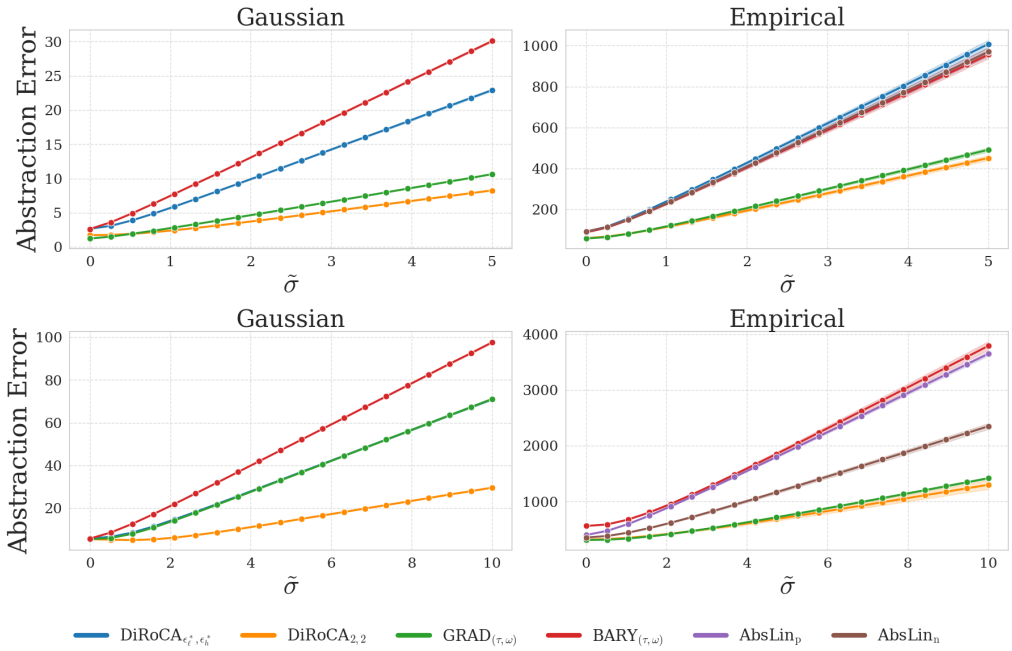


Figure 15: Robustness to *Exponential* noise intensity ($\tilde{\sigma}$) on the **SLC** (top) and **LiLUCAS** (bottom) experiments for the Gaussian (left) and Empirical (right) settings. The evaluation is performed at a fixed outlier fraction of $\alpha = 1.0$ (fully noisy data).

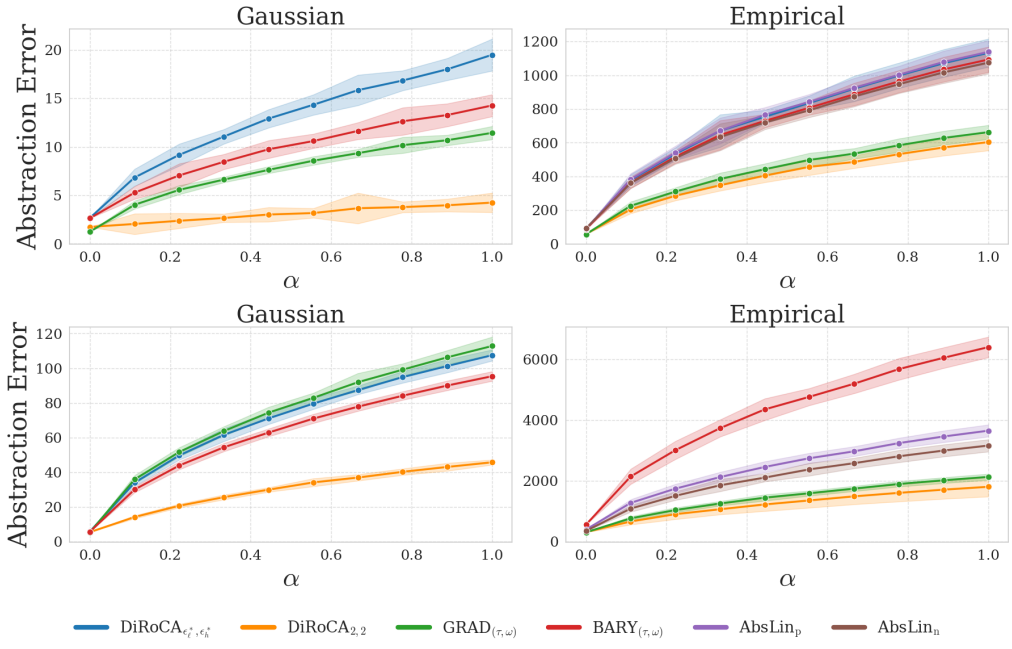


Figure 16: Robustness to outlier fraction (α) on the **SLC** (top) and **LiLUCAS** (bottom) experiments for the Gaussian (left) and Empirical (right) settings. The evaluation is performed at a fixed *Student-t* noise intensity ($\tilde{\sigma} = 5.0$ for SLC and $\tilde{\sigma} = 10.0$ for LiLUCAS)

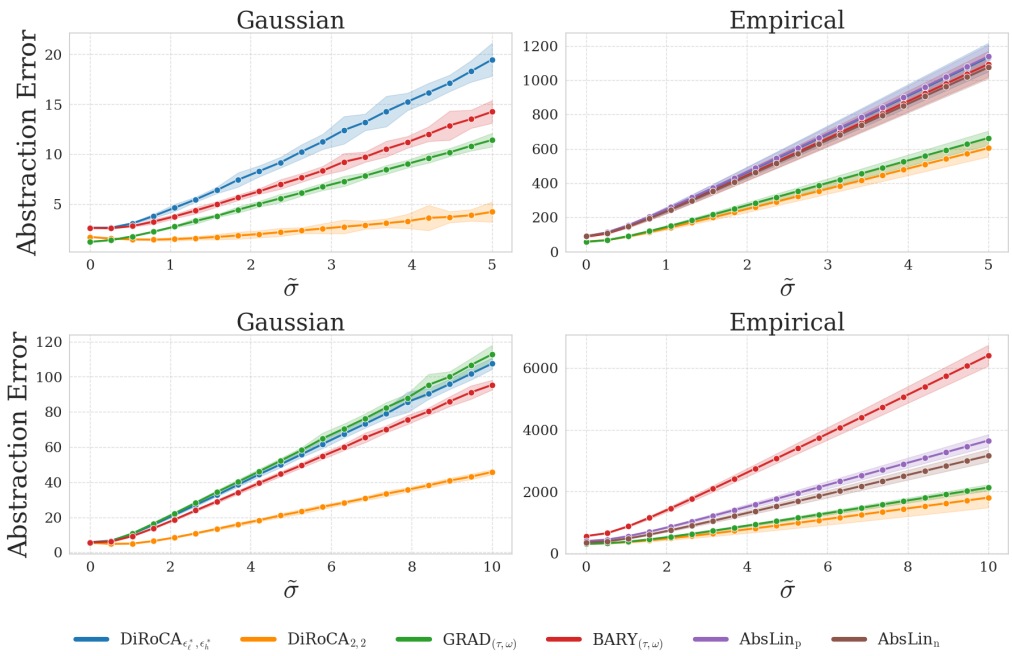


Figure 17: Robustness to *Student-t* noise intensity ($\tilde{\sigma}$) on the **SLC** (top) and **LiLUCAS** (bottom) experiments for the Gaussian (left) and Empirical (right) settings. The evaluation is performed at a fixed outlier fraction of $\alpha = 1.0$ (fully noisy data).

A.8 Optimization

A.8.1 Distributionally Robust Causal Abstraction Learning (DiRoCA)

In this section, we provide a detailed analytical presentation of the optimization procedures of DiRoCA, both for the Gaussian and empirical settings.

Gaussian case. Recall that in the Gaussian case DiRoCA solves the following constrained min-max optimization problem:

$$\min_{\mathrm{T}\in\mathbb{R}^{h\times\ell}} \max_{\substack{\mu^\ell,\Sigma^\ell\\ \mu^h,\Sigma^h}} \mathbb{E}_{\iota\sim q}\Big[\mathcal{W}_2^2\big(\mathcal{N}(\mathrm{T}\,\mathbf{L}_\iota\mu^\ell, \mathrm{T}\,\mathbf{L}_\iota\Sigma^\ell\mathbf{L}_\iota^\top\,\mathrm{T}^\top),\ \mathcal{N}(\mathbf{H}_{\omega(\iota)}\mu^h, \mathbf{H}_{\omega(\iota)}\Sigma^h\mathbf{H}_{\omega(\iota)}^\top)\big)\Big] \tag{48}$$

supported by the environmental uncertainty constraints:

$$\mathcal{W}_2^2(\mathcal{N}(\mu^\ell,\Sigma^\ell),\ \mathcal{N}(\widehat{\mu^\ell},\widehat{\Sigma^\ell})) \leqslant \varepsilon_\ell^2, \quad \mathcal{W}_2^2(\mathcal{N}(\mu^h,\Sigma^h),\ \mathcal{N}(\widehat{\mu^h},\widehat{\Sigma^h})) \leqslant \varepsilon_h^2 \tag{49}$$

where $\widehat{\mu^\ell}, \widehat{\Sigma^\ell}, \widehat{\mu^h}, \widehat{\Sigma^h}$ are empirical estimates from observed data. Since we work with Markovian SCMs, the exogenous variables are jointly independent, which implies that their covariance matrices are diagonal. This structural property simplifies the form of the Wasserstein constraints, allowing them to be expressed as:

$$\|\mu^\ell - \widehat{\mu^\ell}\|_2^2 + \|{\Sigma^\ell}^{1/2} - \widehat{\Sigma^\ell}^{1/2}\|_2^2 \leqslant \varepsilon_\ell^2 \quad \text{and} \quad \|\mu^h - \widehat{\mu^h}\|_2^2 + \|{\Sigma^h}^{1/2} - \widehat{\Sigma^h}^{1/2}\|_2^2 \leqslant \varepsilon_h^2 \tag{50}$$

Expanding the Gelbrich formula for the Gaussian 2-Wasserstein distance, the objective decomposes as:

$$F(\mathrm{T},\mu^\ell,\Sigma^\ell,\mu^h,\Sigma^h) = \underbrace{\mathbb{E}_{\iota\sim q}\Big[\|\,\mathrm{T}\,\mathbf{L}_\iota\mu^\ell - \mathbf{H}_{\omega(\iota)}\mu^h\|_2^2 + \mathrm{Tr}(\mathrm{T}\,\mathbf{L}_\iota\Sigma^\ell\mathbf{L}_\iota^\top\,\mathrm{T}^\top) + \mathrm{Tr}(\mathbf{H}_{\omega(\iota)}\Sigma^h\mathbf{H}_{\omega(\iota)}^\top)\Big]}_{F_s\,\text{(smooth term)}} \tag{51}$$

$$\underbrace{-\mathbb{E}_{\iota\sim q}\Big[2\,\mathrm{Tr}\left((\mathbf{H}_{\omega(\iota)}\Sigma^h\mathbf{H}_{\omega(\iota)}^\top)^{1/2}(\mathrm{T}\,\mathbf{L}_\iota\Sigma^\ell\mathbf{L}_\iota^\top\,\mathrm{T}^\top)(\mathbf{H}_{\omega(\iota)}\Sigma^h\mathbf{H}_{\omega(\iota)}^\top)^{1/2}\right)^{1/2}\Big]}_{F_n\,\text{(non-smooth term)}} \tag{52}$$

We rewrite the objective as the sum of a *smooth* component F_{s} capturing the quadratic terms and a *non-smooth* component F_{n} capturing the last trace term. Thus the objective becomes:

$$\min_{\mathrm{T}} \max_{\mu^\ell, \Sigma^\ell, \mu^h, \Sigma^h} \quad F_{\mathrm{s}}(\mathrm{T}, \mu^\ell, \Sigma^\ell, \mu^h, \Sigma^h) + F_{\mathrm{n}}(\mathrm{T}, \Sigma^\ell, \Sigma^h) \tag{53}$$

Let $S_\ell = \mathrm{T}\,\mathbf{L}_\iota \Sigma^\ell \mathbf{L}_\iota^\top\,\mathrm{T}^\top$ and $S_h = \mathbf{H}_{\omega(\iota)} \Sigma^h \mathbf{H}_{\omega(\iota)}^\top$. Then the non-smooth term simplifies as:

$$\mathrm{Tr}\left(\left(S_h^{1/2} S_\ell S_h^{1/2}\right)^{1/2}\right) = \|S_\ell^{1/2} S_h^{1/2}\|_* \tag{54}$$

To handle the non-smoothness of the nuclear norm in optimization, we use the variational characterization from Srebro et al. [2004]. For any matrix $X = UV^\top$,

$$\|X\|_* = \min_{\substack{U,V: \\ X = UV^\top}} \|U\|_{\mathrm{F}} \|V\|_{\mathrm{F}} \tag{55}$$

In our case:

$$\|S_\ell^{1/2} S_h^{1/2}\|_* \leqslant \|S_\ell^{1/2}\|_{\mathrm{F}} \|S_h^{1/2}\|_{\mathrm{F}} \tag{56}$$

Applying this inequality, we upper bound the non-smooth term, allowing the optimization to proceed efficiently using standard gradient and proximal-gradient methods. Specifically, we maximize a surrogate lower bound $\tilde{F} \leqslant F$:

$$\tilde{F} \coloneqq F_{\mathrm{s}}(\mathrm{T}, \mu^\ell, \Sigma^\ell, \mu^h, \Sigma^h) + F_{\mathrm{n}}^*(\mathrm{T}, \Sigma^\ell, \Sigma^h) \tag{57}$$

where $F_{\mathrm{n}}^*(\mathrm{T}, \Sigma^\ell, \Sigma^h) = -\|(\mathrm{T}\,\mathbf{L}_\iota \Sigma^\ell \mathbf{L}_\iota^\top\,\mathrm{T}^\top)^{1/2}\|_{\mathrm{F}} \|(\mathbf{H}_{\omega(\iota)} \Sigma^h \mathbf{H}_{\omega(\iota)}^\top)^{1/2}\|_{\mathrm{F}}$.

The optimization algorithm alternates between updating T via gradient descent on (51), and updating the moments $(\mu^\ell, \Sigma^\ell), (\mu^h, \Sigma^h)$ via proximal-gradient ascent on (57) (see Algorithm 2). Figures 18a and 18b provide a visual illustration, showing the initial empirical distribution ($\mathcal{N}(\hat{\mu}, \hat{\Sigma})$) and the final worst-case distribution ($\mathcal{N}(\mu^\star, \Sigma^\star)$) learned by the adversary for our two experimental datasets.

Note 1. The proximal operator of the Frobenius norm of a matrix $A \in R^{m \times n}$ is given as

$$\mathrm{prox}_{\lambda \|\cdot\|_F}(A) = \begin{cases} \left(I - \frac{\lambda}{\|A\|_F}\right) A, & \text{if } \|A\|_F > \lambda \\ 0, & \text{otherwise} \end{cases}. \tag{58}$$

Here A denotes the square root of the transformed covariance matrices (e.g.,$(\mathrm{T}\,L_\iota \Sigma^\ell L_\iota^\top\,\mathrm{T}^\top)^{1/2}$ and $(H_{\omega(\iota)} \Sigma^h H_{\omega(\iota)}^\top)^{1/2}$).

Note 2. The projection operation $\mathrm{proj}_{\mathcal{W}_2 \leqslant \varepsilon_h}(\cdot, \cdot, \cdot, \cdot)$ enforces the Wasserstein constraints:

$$\mathcal{W}_2^2\big(\mathcal{N}(\mu, \Sigma),\, \mathcal{N}(\hat{\mu}, \hat{\Sigma})\big) \leqslant \varepsilon^2 \quad \text{via projection onto a Gelbrich ball.} \tag{59}$$

This is done iteratively using the update:

$$(\mu, \Sigma) \leftarrow \hat{\mu} + \alpha(\mu - \hat{\mu}),\, \hat{\Sigma}^{1/2} + \alpha(\Sigma^{1/2} - \hat{\Sigma}^{1/2}) \tag{60}$$

with scaling factor $\alpha = \epsilon / \mathcal{W}_2^2((\mu, \Sigma), (\hat{\mu}, \hat{\Sigma}))$.

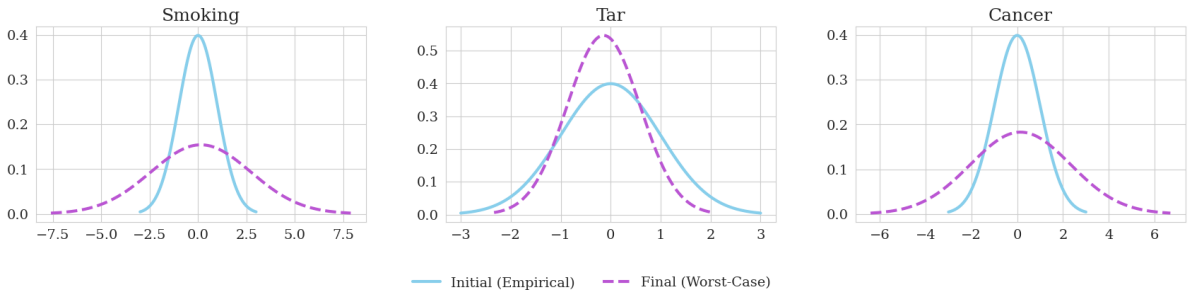
Empirical case. When there are no structural assumptions about the distributional form of the environment, the empirical formulation provides a fully data-driven robust alternative, particularly useful when the low- and high-level environments are available only through finite samples. In this case, we introduce perturbation matrices $\Theta_d \in \mathbb{R}^{N_d \times d}$ over the empirical exogenous samples and define $\rho^d(\Theta_d) = \frac{1}{N_d} \sum_{i=1}^{N_d} \delta_{\hat{u}_i + \theta_i}$. The objective then measures the squared Frobenius discrepancy between the perturbed distributions of the two SCMs:

$$F(\mathrm{T}, \Theta_\ell, \Theta_h) := \mathbb{E}_{\iota \sim q}\left[\left\|\mathrm{T}\;\mathbf{L}_\iota(\mathrm{U}^\ell + \Theta_\ell) - \mathbf{H}_{\omega(\iota)}(\mathrm{U}^h + \Theta_h)\right\|_{\mathrm{F}}^2\right], \tag{61}$$

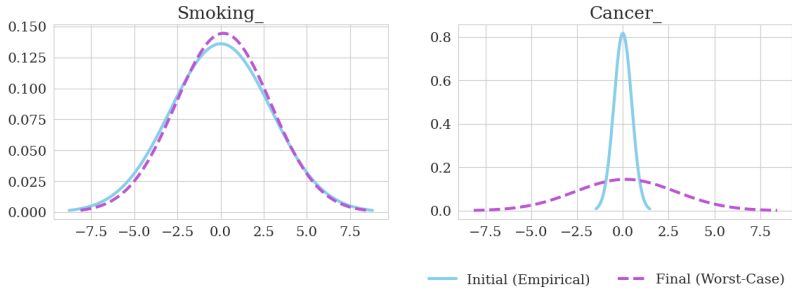
where $\mathrm{U}^\ell, \mathrm{U}^h$ are the empirical exogenous samples for $\mathcal{M}^\ell$ and $\mathcal{M}^h$, respectively. The perturbations capture adversarial shifts around these nominal empirical distributions. A proposed reparameterization reduces the Wasserstein ambiguity sets to constrained Frobenius norm balls [Kuhn et al., 2019]:

$$\|\Theta_\ell\|_F \leqslant \varepsilon_\ell \sqrt{N_\ell}, \quad \|\Theta_h\|_F \leqslant \varepsilon_h \sqrt{N_h} \tag{62}$$

We solve this via alternating projected gradient descent. The optimization alternates between updating the abstraction map T with gradient descent and the adversarial perturbations of the samples through projected gradient ascent. Specifically,



(a) Low-level SCM ($\mathcal{M}^{\ell}$)



(b) High-level SCM ($\mathcal{M}^{h}$)

Figure 18: Marginal noise distributions for each variable in the low-level (top) and high-level (bottom) SCMs of the SLC dataset. In each subplot, the solid blue curve represents the initial empirical noise distribution, and the dashed purple curve shows the corresponding worst-case distribution obtained via the distributionally robust optimization.

- **Minimization step (updating T):** Fix perturbations (Θ_ℓ, Θ_h) applied to the samples and perform gradient descent to update the abstraction map T by minimizing the empirical loss.
- **Maximization step (updating perturbations Θ_ℓ, Θ_h):** Fix T and perform adversarial updates to the perturbations:
  - Perturbations Θ_ℓ are added to U^ℓ and Θ_h to U^h.
  - Gradient ascent steps are performed to maximize the abstraction error.
  - After each ascent step, perturbations are projected onto Frobenius norm balls to ensure they lie within a fixed robustness radius (ε_ℓ for Θ, ε_h for Θ_h), enforcing constraints derived from the Wasserstein DRO ambiguity sets.

A.8.2 Barycentric Abstraction Learning ($\textsc{Bary}_{(\tau,\omega)}$).

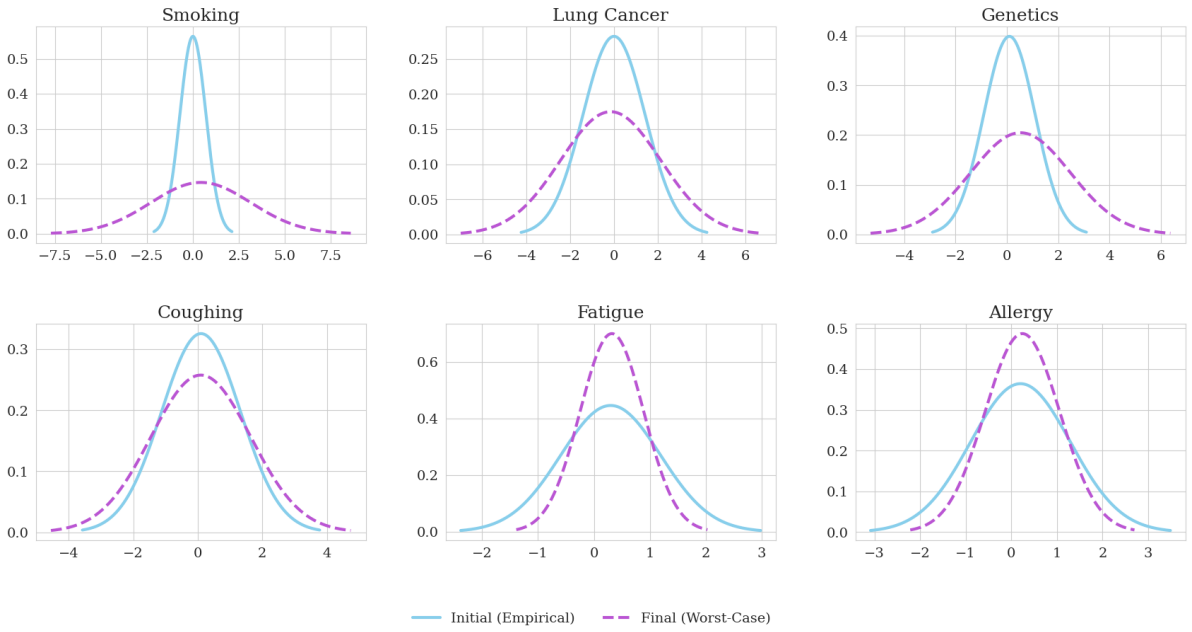
We also consider an abstraction learning method based on *barycentric aggregation* of interventional distributions, inspired by Felekis et al. [2024]. In the case where the environments are Gaussian distributions, the key idea is to aggregate the different interventional distributions at both the low- and high-level SCMs into representative barycentric means and covariances, project the low-level distribution into the high-level space, and then minimize the Wasserstein distance between these two distributions.

Formally, given mixing matrices $\{\mathbf{L}_\iota\}_{\iota\in\mathcal{I}^\ell}$ and $\{\mathbf{H}_\eta\}_{\eta\in\mathcal{I}^h}$ corresponding to the low- and high-level SCMs under interventions, we define the *barycentric mean* at each level according to Eq. 16:

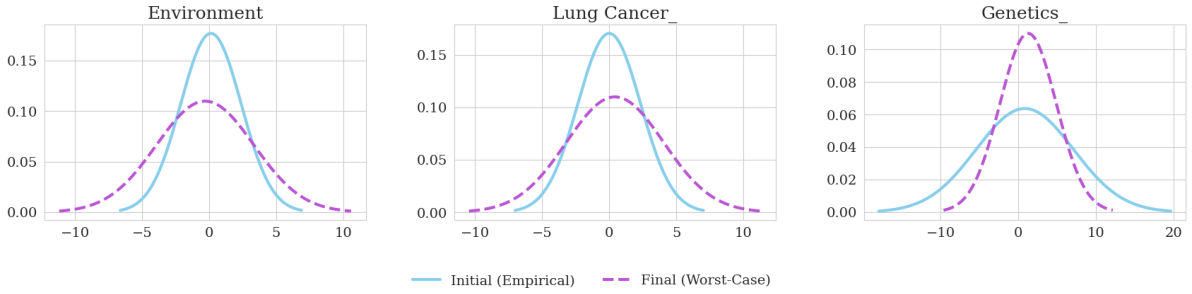
$$\mu^{\ell}_{\text{bary}} = \frac{1}{|\mathcal{I}^\ell|}\sum_{\iota\in\mathcal{I}^\ell}\mathbf{L}_\iota\mu^{\ell}_U, \quad \mu^{h}_{\text{bary}} = \frac{1}{|\mathcal{I}^h|}\sum_{\eta\in\mathcal{I}^h}\mathbf{H}_\eta\mu^{h}_U. \tag{63}$$

The *barycentric covariance* at the low level, $\Sigma^{\ell}_{\text{bary}}$, is obtained as the Wasserstein barycenter of the set $\{\mathbf{L}_\iota\Sigma^{\ell}_U\mathbf{L}_\iota^\top\}_{\iota\in\mathcal{I}^\ell}$, solving the fixed-point equation from Eq. 17:

$$\Sigma^{\ell}_{\text{bary}} = \frac{1}{|\mathcal{I}^\ell|}\sum_{\iota\in\mathcal{I}^\ell}\lambda_\iota\left(\Sigma^{\ell}_{\text{bary}}{}^{1/2}\left(\mathbf{L}_\iota\Sigma^{\ell}_U\mathbf{L}_\iota^\top\right)\Sigma^{\ell}_{\text{bary}}{}^{1/2}\right)^{1/2}, \tag{64}$$



(a) Low-level SCM ($\mathcal{M}^{\ell}$)



(b) High-level SCM ($\mathcal{M}^{h}$)

Figure 19: Marginal noise distributions for each variable in the low-level (top) and high-level (bottom) SCMs of the LiLUCAs dataset. In each subplot, the solid blue curve represents the initial empirical noise distribution, and the dashed purple curve shows the corresponding worst-case distribution obtained via the distributionally robust optimization.

The barycentric covariance Σ^{h}_{bary} at the high level is computed analogously over the set $\{\mathbf{H}_{\eta}\Sigma^{h}_{U}\mathbf{H}_{\eta}^{\top}\}_{\eta\in\mathcal{I}^{h}}$. To map from the low-level to the high-level space, we first project the low-level barycentric distribution onto the high-level dimension. Specifically, letting $V \in \mathbb{R}^{\ell\times h}$ denote the projection matrix, obtained via PCA, we define the projected mean and covariance:

$$\mu^{\ell}_{\text{proj}} = V^{\top}\mu^{\ell}_{\text{bary}}, \quad \Sigma^{\ell}_{\text{proj}} = V^{\top}\Sigma^{\ell}_{\text{bary}}V. \tag{65}$$

Consequently, by defining the abstraction map as the optimal transport Monge map between the distributions $\mathcal{N}(\mu^{\ell}_{\text{proj}}, \Sigma^{\ell}_{\text{proj}})$ and $\mathcal{N}(\mu^{h}_{\text{bary}}, \Sigma^{h}_{\text{bary}})$, from Eq. 12, since both distributions are Gaussian, this admits the following closed-form expression:

$$T(x) = \mu^{h}_{\text{bary}} + A(x - \mu^{\ell}_{\text{proj}}), \quad \text{where} \quad A = {\Sigma^{h}_{\text{bary}}}^{1/2}\left(\Sigma^{\ell}_{\text{proj}}\right)^{-1/2}. \tag{66}$$

Finally, since we initially projected the low-level distribution, the full transformation matrix is:

$$\mathrm{T} = AV^{\top}, \tag{67}$$

In both cases, the resulting abstraction approximately aligns the barycentric summaries of the low- and high-level interventional distributions.

In the setting where only empirical samples of the environments are available and no structural assumptions are made about their underlying distributions, we simply take the arithmetic mean across all the interventions of the mixing matrices $\mathbf{L}_{bary}$ and $\mathbf{H}_{bary}$ and propagate the exogenous datasets U^{ℓ} and U^{h} through them and solve with gradient descent the following problem:

$$\min_{\mathrm{T}\in\mathbb{R}^{h\times\ell}} \left\| \mathrm{T}\ \mathbf{L}_{bary}(\mathrm{U}^{\ell}) - \mathbf{H}_{bary}(\mathrm{U}^{h}) \right\|_{\mathrm{F}}^{2} \tag{68}$$

While $\textsc{Bary}_{(\tau,\omega)}$ is computationally efficient, it does not explicitly optimize robustness to distributional shifts, unlike DIROCA.

Algorithm 2 Gaussian DiRoCA

Input: $\mathcal{S}^\ell, \mathcal{S}^h, \mathcal{I}^\ell, \mathcal{I}^h, \omega : \mathcal{I}^\ell \to \mathcal{I}^h$, samples from $\mathbb{P}_{\mathcal{M}^\ell_\iota}(\mathcal{X}^\ell), \mathbb{P}_{\mathcal{M}^h_{\omega(\iota)}}(\mathcal{X}^h)$
$\widehat{\rho^\ell} \sim \mathcal{N}(\widehat{\mu}_\ell, \widehat{\Sigma_\ell}) \leftarrow (\mathbf{g}^\ell_\#(\mathbb{P}_{\mathcal{M}^\ell}))^{-1}$ ▷ Abduction for $\mathcal{M}^\ell$
$\widehat{\rho^h} \sim \mathcal{N}(\widehat{\mu_h}, \widehat{\Sigma_h}) \leftarrow (\mathbf{g}^h_\#(\mathbb{P}_{\mathcal{M}^h}))^{-1}$ ▷ Abduction for $\mathcal{M}^h$
$\epsilon \leftarrow$ from Theorem 1 ▷ Compute robustness radius
$\mathbb{B}_{\epsilon,2}(\widehat{\boldsymbol{\rho}}) \leftarrow \mathbb{B}_{\varepsilon_\ell,2}(\widehat{\rho^\ell}) \times \mathbb{B}_{\varepsilon_h,2}(\widehat{\rho^h})$
Initialize: $\mathrm{T}^{(0)} \in \mathbb{R}^{h\times\ell}, \boldsymbol{\rho}^{(0)} = \mathcal{N}(\mu^{\ell(0)}, \Sigma^{\ell(0)}) \otimes \mathcal{N}(\mu^{h(0)}, \Sigma^{h(0)})$, $\eta_{\mathrm{T}}, \eta_{\boldsymbol{\rho}}$, $k_{\min}$, $k_{\max}$
repeat ▷ min-max iterations
  for $k_{\min}$ steps **do** ▷ Gradient descent: *Update abstraction* T
    $\mathrm{T}^{(t+1)} \leftarrow \mathrm{T}^{(t)} - \eta_{\mathrm{T}} \cdot \nabla_{\mathrm{T}} F(\mathrm{T}^{(t)}, \rho^{(t)})$
  end for
  for $k_{\max}$ steps **do** ▷ Proximal-gradient Ascent: *Update joint environment* $\boldsymbol{\rho}$
    $\mu^{\ell(t+1)} \leftarrow \mu^{\ell(t)} + \eta_{\boldsymbol{\rho}} \cdot \nabla_{\mu^\ell} F(\mu^\ell, \mathrm{T}^{(t+1)})$
    $\mu^{h(t+1)} \leftarrow \mu^{h(t)} + \eta_{\boldsymbol{\rho}} \cdot \nabla_{\mu^h} F(\mu^h, \mathrm{T}^{(t+1)})$
    $\Sigma^{\ell(t+\frac{1}{2})} \leftarrow \Sigma^{\ell(t)} + \eta_\rho \cdot \nabla_{\Sigma^\ell} \tilde{F}(\tau^{(t+1)}, \rho^{(t)})$
    $\Sigma^{\ell(t+1)} \leftarrow \mathrm{prox}_{\lambda_\ell}(\Sigma^{\ell(t+\frac{1}{2})})$ ▷ Proximal step
    $\Sigma^{h(t+\frac{1}{2})} \leftarrow \Sigma^{h(t)} + \eta_\rho \cdot \nabla_{\Sigma^h} \tilde{F}(\tau^{(t+1)}, \rho^{(t)})$
    $\Sigma^{h(t+1)} \leftarrow \mathrm{prox}_{\lambda_h}(\Sigma^{h(t+\frac{1}{2})})$ ▷ Proximal step
    $(\mu^{\ell(t+1)}, \Sigma^{\ell(t+1)}) \leftarrow \mathrm{proj}_{\mathcal{W}_2 \leqslant \varepsilon_\ell}(\mu^{\ell(t+1)}, \Sigma^{\ell(t+1)}, \widehat{\mu^\ell}, \widehat{\Sigma^\ell})$ ▷ Project to Gelbrich ball
    $(\mu^{h(t+1)}, \Sigma^{h(t+1)}) \leftarrow \mathrm{proj}_{\mathcal{W}_2 \leqslant \varepsilon_h}(\mu^{h(t+1)}, \Sigma^{h(t+1)}, \widehat{\mu^h}, \widehat{\Sigma^h})$ ▷ Project to Gelbrich ball
    $\boldsymbol{\rho}^{(t+1)} \leftarrow \mathcal{N}(\mu^{\ell(t+1)}, \Sigma^{\ell(t+1)}) \otimes \mathcal{N}(\mu^{h(t+1)}, \Sigma^{h(t+1)})$
  end for
convergence
return $\mathrm{T}^\star \in \mathbb{R}^{h\times\ell}, \boldsymbol{\rho}^\star = \rho^{\star\ell} \otimes \rho^{\star h}$

Algorithm 3 Empirical DiRoCA

Input: $\mathcal{S}^\ell, \mathcal{S}^h, \mathcal{I}^\ell, \mathcal{I}^h, \omega : \mathcal{I}^\ell \to \mathcal{I}^h$, samples from $\mathbb{P}_{\mathcal{M}^\ell_\iota}(\mathcal{X}^\ell), \mathbb{P}_{\mathcal{M}^h_{\omega(\iota)}}(\mathcal{X}^h)$
$\widehat{\rho^\ell} = \frac{1}{N_\ell}\sum_{i=1}^{N_\ell} \delta_{\widehat{u^\ell}_i} \leftarrow \mathbf{L}^{-1}\mathcal{X}^\ell$ ▷ Abduction for $\mathcal{M}^\ell$
$\widehat{\rho^h} = \frac{1}{N_h}\sum_{i=1}^{N_h} \delta_{\widehat{u^h}_i} \leftarrow \mathbf{H}^{-1}\mathcal{X}^h$ ▷ Abduction for $\mathcal{M}^h$
$\widehat{\boldsymbol{\rho}} \leftarrow \widehat{\rho^\ell} \otimes \widehat{\rho^h}$ ▷ Joint environment
$\epsilon \leftarrow$ from Theorem 2 ▷ Robustness radius
Initialize: $\mathrm{T}^{(0)}, \Theta_\ell^{(0)}, \Theta_h^{(0)}$
$\rho^{\ell(0)} = \frac{1}{N_\ell}\sum_{i=1}^{N_\ell} \delta_{\widehat{u}_i^\ell + \theta_i^{\ell(0)}}$
$\rho^{h(0)} = \frac{1}{N_h}\sum_{i=1}^{N_h} \delta_{\widehat{u}_i^h + \theta_i^{h(0)}}$
$\boldsymbol{\rho}^{(0)} \leftarrow \rho^{\ell(0)} \otimes \rho^{h(0)}$
repeat ▷ min-max iterations
  for $k_{\min}$ steps **do** ▷ Gradient descent: *Update abstraction* T
    $\mathrm{T}^{(t+1)} \leftarrow \mathrm{T}^{(t)} - \eta \nabla_{\mathrm{T}} F(\mathrm{T}^{(t)}, \Theta_\ell^{(t)}, \Theta_h^{(t)})$
  end for
  for $k_{\max}$ steps **do** ▷ Gradient Ascent: *Update joint environment* $\boldsymbol{\rho}$
    $\Theta_\ell^{\text{temp}} \leftarrow \Theta_\ell^{(t)} + \eta \nabla_{\Theta_\ell} F(\mathrm{T}^{(t+1)}, \Theta_\ell^{(t)}, \Theta_h^{(t)})$
    $\Theta_\ell^{(t+1)} \leftarrow \text{proj}_{\|\cdot\|_F \leqslant \varepsilon_\ell \sqrt{N_\ell}}(\Theta_\ell^{\text{temp}})$ ▷ Project onto Frobenious norm ball
    $\Theta_h^{\text{temp}} \leftarrow \Theta_h^{(t)} + \eta \nabla_{\Theta_h} F(\mathrm{T}^{(t+1)}, \Theta_\ell^{(t)}, \Theta_h^{(t)})$
    $\Theta_h^{(t+1)} \leftarrow \text{proj}_{\|\cdot\|_F \leqslant \varepsilon_h \sqrt{N_h}}(\Theta_h^{\text{temp}})$ ▷ Project onto Frobenious norm ball
    $\rho^{\ell(t+1)} \leftarrow \frac{1}{N_\ell}\sum_{i=1}^{N_\ell} \delta_{\widehat{u}_i^\ell + \theta_i^{\ell(t+1)}}$
    $\rho^{h(t+1)} \leftarrow \frac{1}{N_h}\sum_{i=1}^{N_h} \delta_{\widehat{u}_i^h + \theta_i^{h(t+1)}}$
    $\boldsymbol{\rho}^{(t+1)} \leftarrow \rho^{\ell(t+1)} \otimes \rho^{h(t+1)}$
  end for
convergence
return $\mathrm{T}^\star \in \mathbb{R}^{h \times \ell}, \boldsymbol{\rho}^\star$